



# **Towards Emotionally Intelligent Task-oriented Dialogue Systems**

Inaugural-Dissertation

zur Erlangung des Doktorgrades  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

**Shutong Feng**  
aus China

Düsseldorf, March 2026

aus dem Institut für Informatik  
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der  
Mathematisch-Naturwissenschaftlichen Fakultät der  
Heinrich-Heine-Universität Düsseldorf

Berichtersteller:

1. Prof. Dr. Milica Gašić
2. Prof. Dr. Ryuichiro Higashinaka

Tag der mündlichen Prüfung: 4. February 2026

# Declaration of Authorship

I, *Shutong Feng*, declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf.

Düsseldorf,

---

Location, Date

---

Shutong Feng



# Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Milica Gašić, whose guidance, support, and unwavering encouragement have shaped every stage of this PhD journey. Your insight, patience, and high standards have continually pushed me to grow as a researcher.

I would also like to thank my colleagues, Dr. Nurul Lubis, Dr. Michael Heck, Dr. Benjamin Ruppik, Dr. Christian Geishauser, Dr. Hsien-chin Lin, Dr. Carel van Niekerk, and Renato Vukovic, for creating an inspiring and supportive environment, as well as for their collaborative efforts and assistance with proofreading across various projects. In particular, I am especially grateful to my co-supervisor, Dr. Nurul Lubis, whose help was invaluable, especially during the early stages of this journey. Your mentorship provided a foundation that I continued to build upon throughout the project. I would also like to thank our group's secretaries, Annette Peitsch and Verena Schlösser-Lewin, for their kind and reliable assistance.

To my friends, thank you for bringing laughter and courage into my life. Your encouragement made this long journey far more manageable and meaningful.

Last but certainly not least, I owe my greatest thanks to my mother, Xiuli Zhang and my father, Peihai Feng. Your unconditional love, support, and belief in me have carried me through the most challenging moments. This achievement would not have been possible without everything you have given me.



# Abstract

The development of conversational agents has long been driven by the ambition to create systems that interact with humans in ways that are both functional and emotionally attuned. While task-oriented dialogue (ToD) systems are designed to achieve specific goals such as booking a restaurant or suggesting a tourist attraction, they typically overlook the emotional dimension of human communication. Yet, beyond transmitting factual information, interlocutors continually signal and interpret intentions through emotions. These signals influence error recovery and user satisfaction, all of which are indispensable in human communications and crucial to the system’s functionality. The omission of emotions in ToD systems, while keeping the system centred on functional goals, limits user satisfaction and reduces robustness in real-world interactions.

In this thesis, we investigate how emotion can be integrated into ToD systems to enhance their effectiveness. The work unfolds through a sequence of contributions that move from the preparation of resources to the development of robust modelling methods, and finally to the incorporation of emotion into the full ToD pipeline.

First, we address the scarcity of suitable resources by introducing *EmoWOZ*, the first large-scale corpus of ToDs annotated with a dedicated user emotion taxonomy designed to capture the subtle affective behaviours unique to ToDs. *EmoWOZ* contains both human-human and human-machine conversations covering a wide spectrum of user emotions. Annotated with high-quality labels validated through rigorous quality control, *EmoWOZ* provides a foundation for studying user emotion in ToDs and supports downstream ToD modelling tasks.

Second, we explore methods for modelling emotion in ToDs, pursuing two complementary directions. On the one hand, we develop lightweight supervised models by adapting chit-chat emotion recognition in conversations (ERC) models to the task-oriented setting. Our framework tackles the challenge from three angles - data, features, and objectives - through targeted data augmentation, the integration of task-specific and emotion-aware features, and novel optimisation objectives that account for relationships among emotion labels. On the other hand, we examine the potential of large language models (LLMs) as flexible, general-purpose emotion recognisers. We evaluate them across multiple dialogue settings, including ToDs, chit-chat, and psychological interviews, and test their effectiveness under low-resource scenarios and speech recognition errors. Together, these investigations establish a complementary toolkit for emotion recognition: specialised supervised models offer accuracy and efficiency in well-defined ToD settings, whereas LLMs provide flexibility and robustness when dealing with broader emotion taxonomies and less constrained dialogue scenarios.

Finally, we integrate emotions into the full ToD pipeline and systematically investigate optimal design choices for emotionally intelligent ToD systems. In our first line of work, we focus on the practical challenges to infuse emotion-awareness into both modular and end-to-end ToD systems. This is achieved by extending the dialogue state to include user emotion, expanding the dialogue policy to include affective actions, and conditioning the natural language generation both semantically and emotionally. For the modular system, we further employ multi-objective online reinforcement learning (RL) to optimise task success and emotional appropriateness jointly. Evaluations with simulated and real users demonstrate that emotion-aware systems not only improve task performance but also enrich the user’s emotional experience, confirming the practical benefits of incorporating emotion into ToDs. Building on these insights, our second line of work systematically explores a wider range of design considerations spanning architecture, representation, and optimisation. We present *LUSTER* (LLM-based Unified System for Task-oriented dialogue with End-to-end Reinforcement learning). *LUSTER* achieves significantly improved task efficiency and user satisfaction by leveraging LLM-based end-to-end architecture, fully lexicalised representations, and online multi-objective RL optimisation.

Together, the advances discussed in this thesis demonstrate that emotion can be seamlessly integrated into ToD systems, not only enhancing their effectiveness and robustness but also bringing them closer to the goal of human-like, emotionally intelligent interaction.



# Contents

<b>List of Figures</b>	<b>xiv</b>
<b>Summary of Notation</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Task-oriented Dialogue Systems . . . . .	2
1.2 Emotionally Intelligent Task-oriented Dialogue Agent . . . . .	4
Role of Emotion in Task-oriented Dialogues . . . . .	5
Key Abilities of Emotional Intelligence . . . . .	5
Multimodality of Emotion . . . . .	6
1.3 Challenges in Building Emotionally Intelligent Task-oriented Dialogue Systems . . . . .	7
1.3.1 Data Scarcity and Annotation Challenges . . . . .	7
1.3.2 Integration of Emotion into Task-oriented Dialogue Systems . . . . .	8
Emotion Perception . . . . .	8
Emotion Processing . . . . .	8
Emotion Expression . . . . .	8
1.4 Contributions . . . . .	9
1.4.1 Dataset Construction for Emotion Modelling in Task-oriented Dialogues . . . . .	9
1.4.2 Emotion Recognition Learning in Task-oriented Dialogues . . . . .	9
1.4.3 Integrating Emotion into Task-oriented Dialogue Systems . . . . .	10
1.5 Thesis Structure . . . . .	11
<b>2 Learning from Data</b>	<b>13</b>
2.1 Overview . . . . .	13
2.2 Data Collection . . . . .	14
2.2.1 Source of Data . . . . .	14
2.2.2 Annotation and Labelling . . . . .	14
Inter-annotator Agreement . . . . .	15
Interpretation of Inter-annotator Agreement Metrics . . . . .	16
2.2.3 Ethical and Practical Considerations . . . . .	16
2.3 Neural Network Function Approximator . . . . .	17
2.3.1 Neural Networks as Universal Function Approximators . . . . .	17
2.3.2 Non-linear Activation Functions . . . . .	18
2.3.3 Representation of Text in Neural Models . . . . .	20
2.4 Neural Network Optimisation . . . . .	20
2.4.1 Loss Functions . . . . .	21
Cross-entropy Loss . . . . .	21
Mean Squared Error Loss . . . . .	21
Custom Losses . . . . .	21
2.4.2 Optimisers . . . . .	21
Gradient Descent and Stochastic Gradient Descent . . . . .	22
Adam and AdamW . . . . .	22
2.4.3 Practical Optimisation Techniques . . . . .	23
Normalisation . . . . .	23

	Gradient Clipping . . . . .	23
	Regularisation . . . . .	24
2.5	Neural Network Architectures . . . . .	24
2.5.1	Feed-forward Neural Networks . . . . .	24
	Gradient Vanishing and Exploding . . . . .	25
2.5.2	Residual Neural Networks . . . . .	26
2.5.3	Recurrent Neural Networks . . . . .	26
	The Vanishing Gradient Problem . . . . .	27
	Encoder-decoder Architecture of RNN . . . . .	28
	Attention Mechanisms in RNNs . . . . .	28
2.5.4	Transformers . . . . .	29
	Attention Mechanism in Transformers . . . . .	29
	Multi-Head Attention (MHA) . . . . .	30
	Transformer Block . . . . .	30
	Positional Encoding . . . . .	32
	The Encoder-decoder Architecture of Transformers . . . . .	32
2.6	Supervised Learning . . . . .	33
2.6.1	Formal Definition . . . . .	34
2.6.2	Classification . . . . .	35
2.6.3	Regression . . . . .	35
2.6.4	Sequence-to-Sequence Learning . . . . .	35
2.6.5	Advantages and Limitations . . . . .	36
2.7	Transfer Learning . . . . .	36
2.7.1	Domain Adaptation . . . . .	37
2.7.2	Data Augmentation . . . . .	37
2.7.3	Multitask-Learning . . . . .	37
2.7.4	Parameter Efficient Fine-tuning . . . . .	37
2.7.5	Advantages and Limitations . . . . .	38
2.8	Unsupervised Learning . . . . .	39
2.8.1	Language Modelling . . . . .	39
	Autoregressive Language Models . . . . .	39
	Masked Language Models . . . . .	39
2.8.2	Advantages and Limitations . . . . .	40
2.9	Semi-Supervised Learning . . . . .	40
2.9.1	Semi-Supervised Learning Techniques . . . . .	40
2.9.2	Advantages and Limitations . . . . .	41
2.10	Reinforcement Learning . . . . .	41
2.10.1	Formal Definition . . . . .	41
2.10.2	Return and Value Functions . . . . .	42
2.10.3	Estimation of Value Functions . . . . .	42
2.10.4	Function Estimation in Deep Reinforcement Learning . . . . .	43
2.10.5	Policy Optimisation . . . . .	43
	Value-based and Policy-based Approaches . . . . .	43
	Online and Offline Reinforcement Learning . . . . .	44
	On-policy and Off-policy Approaches . . . . .	44
2.10.6	Advantages and Limitations . . . . .	44
2.11	Large Language Models . . . . .	44
2.11.1	Scaling Laws . . . . .	45
2.11.2	Training and Alignment . . . . .	45
2.11.3	Emergent Abilities . . . . .	46
2.11.4	Advantages and Limitations . . . . .	47

2.12	Conclusion . . . . .	47
<b>3</b>	<b>Affective Computing</b>	<b>49</b>
3.1	Overview . . . . .	49
3.2	Taxonomy of Emotion . . . . .	50
3.2.1	Dimensional Models of Emotion . . . . .	50
	Two-dimensional Models . . . . .	50
	Three-dimension Models . . . . .	52
	Strengths and Limitations of Dimensional Models . . . . .	52
3.2.2	Categorical Models of Emotion . . . . .	53
	The Theory of Basic Emotions . . . . .	53
	Expanded Taxonomies of Emotions . . . . .	53
	Data-Driven Emotion Taxonomies . . . . .	54
	Strengths and Limitations of Categorical Models . . . . .	54
3.2.3	Hybrid Models of Emotion . . . . .	55
	Plutchik’s Wheel of Emotions . . . . .	55
	The Ortony, Clore, and Collins Model . . . . .	55
	Strengths and Limitations of Hybrid Models . . . . .	56
3.3	Emotion Intelligence in Computers . . . . .	56
3.3.1	Emotion Perception . . . . .	58
3.3.2	Emotion Expression . . . . .	58
3.3.3	Emotion Processing . . . . .	59
3.4	Conclusion . . . . .	59
<b>4</b>	<b>Dialogue Theory</b>	<b>61</b>
4.1	Overview . . . . .	61
4.2	Core Theories of Dialogue . . . . .	61
4.2.1	Speech Act Theory . . . . .	62
4.2.2	The Cooperative Principle . . . . .	62
4.2.3	Grounding in Conversations . . . . .	63
4.2.4	Partially Observable Markov Decision Process . . . . .	64
4.3	Dialogue Settings . . . . .	66
4.3.1	Task-oriented Dialogues . . . . .	66
4.3.2	Chit-chat Dialogues . . . . .	67
4.3.3	Comparative Perspective . . . . .	68
4.3.4	Other Dialogue Settings . . . . .	69
4.4	Conclusion . . . . .	70
<b>5</b>	<b>EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems</b>	<b>71</b>
5.1	Summary . . . . .	71
5.2	Personal Contributions . . . . .	71
<b>6</b>	<b>From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue</b>	<b>91</b>
6.1	Summary . . . . .	91
6.2	Personal Contributions . . . . .	91
<b>7</b>	<b>Affect Recognition in Conversations Using Large Language Models</b>	<b>111</b>
7.1	Summary . . . . .	111
7.2	Personal Contributions . . . . .	111

<b>8</b>	<b>Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation</b>	<b>127</b>
8.1	Summary . . . . .	127
8.2	Personal Contributions . . . . .	127
<b>9</b>	<b>Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation</b>	<b>149</b>
9.1	Summary . . . . .	149
9.2	Personal Contributions . . . . .	149
<b>10</b>	<b>Conclusion</b>	<b>169</b>
10.1	Summary of Key Findings . . . . .	169
10.2	Limitations . . . . .	170
10.3	Recommendations for Future Research . . . . .	170
	<b>Bibliography</b>	<b>173</b>

# List of Figures

1.1	The ontology defines the scope of what a ToD system can understand and talk about.	2
1.2	The ToD system constructs the dialogue state to capture the dialogue context and the user’s goals and constraints. This state is subsequently used to form database queries that return entities fulfilling the user’s search constraints. . . . .	3
1.3	A ToD system interacts with the user in a noisy environment while querying the database symbolically. Its primary objective is task success, but it also aims to enhance the user’s emotional experience by accurately conveying factual knowledge with appropriate sentiment. . . . .	4
1.4	ToD systems follow either a modular or an end-to-end architecture, employing key modelling stages of understanding, management, and generation. . . . .	5
1.5	Capabilities of emotional intelligence can be mapped alongside the ToD modelling pipeline and formulated as computational tasks. . . . .	6
1.6	Illustration of a ToD (hotel booking) versus a chit-chat dialogue (talking about the booked hotel). In ToD, emotions are conveyed implicitly and tied to task progress (e.g., urgency, relief), whereas in chit-chat they are expressed more openly, reflecting personal experiences and social sharing. . . . .	7
2.1	Illustrations of sigmoid and hyperbolic tangent activation functions. . . . .	18
2.2	Illustrations of the ReLU activation function and its variants. . . . .	19
2.3	An illustration of a FNN architecture with input $x$ and output $\hat{y}$ . Each layer $l$ is parameterised with a trainable weight matrix $W_l$ and a bias vector $b_l$ , and has a non-linear activation function $\phi_l$ . . . . .	25
2.4	An illustration of a residual connection. The input $x$ skips and adds to the output of the non-linear residual block $f_{\text{resnet}}(x; \mathbf{W}, \mathbf{b})$ to output the hidden activation $h^{(1)}$ . . . .	26
2.5	An illustration of a recurrent neural network architecture with the unfolded view to show the processing of each time step (token in the sequence). Each RNN cell, $f_{\text{RNN}}(x_t, \mathbf{h}_{t-1}; \theta)$ , represents the operation specified in Equations 2.44 and 2.45. . . . .	27
2.6	An illustration of an encoder-decoder recurrent neural network architecture. The hidden state of the last input token $h_N$ is used as the context vector, which is used as the first hidden state $s_0$ in the decoder to initialise the generation. . . . .	28
2.7	An illustration of an implementation of the attention mechanism in an encoder-decoder recurrent neural network architecture. The attention scoring function in Equation 2.50 is implemented as $f_{\text{score}}(\mathbf{h}_n, \mathbf{s}_m; \theta_{\text{score}}) = \tanh(\mathbf{W}_x \mathbf{h}_n + \mathbf{W}_y \mathbf{s}_m + \mathbf{b}_{\text{score}})$ where trainable parameters $\theta_{\text{score}}$ are weight matrices $\mathbf{W}_x, \mathbf{W}_y$ and bias vector $\mathbf{b}_{\text{score}}$ . Attention scores of all encoder hidden states are normalised with the softmax function, following a weighted sum operation to obtain the context vector $c_m$ (Equation 2.52). In the illustration, $m = 1$ . . . . .	29
2.8	An illustration of the scaled dot-product attention in a multi-head attention set-up. . . .	30

2.9	Left: Transformer block with self-attention takes a sequence of vectors as input stacked in a matrix $X$ and outputs a sequence of hidden vectors stacked in a matrix $H$ . The Transformer block starts with a multi-head attention (MHA), which takes $X$ as queries, keys and values, followed by a residual connection (Add operation) and layer normalisation (Norm). The output of MHA is fed into a FNN with a non-linear activation function. Right: the Transformer block has an additional cross-attention mechanism, which takes an additional sequence of vectors stacked in a matrix $Z$ as keys and values while taking the queries from the output of the previous self-attention module. . . . .	31
2.10	A visualisation of positional encoding values across different positions and dimensions.	33
2.11	An illustration of the Transformer-encoder-decoder architecture. Left: the Transformer encoder is composed of a stack of $L_{\text{enc}}$ Transformer blocks with self-attention. Right: the Transformer decoder is composed of a stack of $L_{\text{dec}}$ Transformer blocks with self-attention and cross-attention, followed by a linear transformation and a softmax activation function, and auto-regressively produces the next element. The positional encodings are added to the input for both the encoder and the decoder. . . . .	34
2.12	An illustration of the interaction between an RL agent and the environment. In each time step, the agent takes action $A_t$ based on the observed state $S_t$ from the environment. The agent then receives the reward $R_{t+1}$ and the next state $S_{t+1}$ from the environment. . . . .	41
2.13	An illustration of (a) online reinforcement learning, (b) off-policy reinforcement learning, and (c) offline reinforcement learning. . . . .	43
2.14	An illustration of Reinforcement Learning from Human Feedback (RLHF) for training a large language model (LLM). It follows three steps: (1) human demonstrations are collected and they are used to initialise the policy $\pi_{\theta, \text{SFT}}$ (the LLM) via supervised fine-tuning (SFT) using the autoregressive language modelling loss in Equation 2.83. (2) For each prompt $x$ , outputs $y^{(1)}, \dots, y^{(k)}$ are sampled from the policy. Human annotators rank these sampled outputs to create a preference dataset, which is used to train the reward model $R_\phi$ , following the loss in Equation 2.100. (3) For each new prompt $x$ , the policy generates an output $y$ . The reward model assigns a reward $r$ for reinforcement learning of the policy following the gradient in Equation 2.102. . . . .	46
3.1	An illustration of (a) the Circumplex Model of Affect, and (b) the positive activation-negative activation (PANA) model. Conceptually, PANA can be viewed as a 45-degree rotation of the Circumplex Model of Affect. . . . .	51
3.2	An illustration of the Vector Model of emotion. . . . .	52
3.3	An illustration of the Valence-Arousal-Dominance model of emotions. . . . .	53
3.4	An illustration of the expansion of categorical taxonomies: from emphasising universality, to considering context and appraisal, and to observing from data. . . . .	54
3.5	An illustration of Plutchik's wheel of emotions (Source: Wikimedia Commons, "Plutchik-wheel.svg"; License: Public Domain). . . . .	55
3.6	An illustration of the OCC model of emotions. In the OCC model, emotions are defined as valenced reactions to elicitors: consequences of events, actions of agents, and aspects of objects. The intensity depends on the attributes of the elicitor. . . . .	57
4.1	A comparison of dialogue settings across theoretical, collaborative, and computational dimensions, with emphasis on the role of emotion. . . . .	68

# Summary of Notation

This chapter summarises the notations used throughout this thesis for quick reference.

## General Mathematical Notations

$a, b, c$	Scalars
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	Vectors
$a_i$	$i$ -th element of vector $\mathbf{a}$ .
$\mathbf{A}, \mathbf{B}, \mathbf{C}$	Matrices
$\mathbf{a}_j$	$j$ -th row of matrix $\mathbf{A}$ .
$\mathbf{a} = [a_i]_{i=1}^I$	Vector $\mathbf{a}$ composed of elements $a_i$ .
$\mathbf{A} = [\mathbf{a}_i]_{i=1}^I$	Matrix $\mathbf{A}$ composed of rows $\mathbf{a}_i$ .
$a_{i,j}$	Element in the $i$ -th row and $j$ -th column of matrix $\mathbf{A}$
$\mathbf{A}^\top$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
$\mathbf{AB}$	Matrix multiplication of $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{x}_{1:t} = \langle \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t \rangle$	Sequence of $t$ vectors.
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	Sets of elements.
$x \in \mathcal{X}$	$x$ is an element of set $\mathcal{X}$ .
$\mathbf{a} \odot \mathbf{b}$	Element-wise multiplication of vectors $\mathbf{a}$ and $\mathbf{b}$ .
$\mathbf{a} \cdot \mathbf{b}$	Dot product of vectors $\mathbf{a}$ and $\mathbf{b}$ .
$\mathbf{a} \oplus \mathbf{b}$	Concatenation of vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a}^2 = \mathbf{a} \odot \mathbf{a}$	Element-wise square of vector $\mathbf{a}$ .
$f(x), g(x)$	Functions dependent on variable $x$
$f(\mathbf{x}; \boldsymbol{\theta})$	Function of $\mathbf{x}$ parameterised by $\boldsymbol{\theta}$
$f \circ g(x) = f(g(x))$	Composition of functions $f$ and $g$
$f: \mathcal{X} \rightarrow \mathcal{Y}$	Function mapping elements from set $\mathcal{X}$ to set $\mathcal{Y}$ .
$\int f(x)dx$	Integral of $f(x)$
$\mathbb{R}$	The set of real numbers
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}^d$	$d$ -dimensional Euclidean space
$\mathbb{R}^{d_1 \times d_2}$	Real-valued matrices with $d_1$ rows and $d_2$ columns
$\Sigma$	Summation symbol
$\log(x)$	Natural logarithm of $x$
$e^x = \exp(x)$	Exponential function of $x$
$\arg \max_y f(y)$	Value of $y$ that maximises the function $f(y)$
$\max_y f(y)$	Maximum value of function $f(y)$ with respect to $y$
$\mathcal{O}(n)$	Big-O notation, indicating the computational complexity of a function

## Notations in Machine Learning

$\mathbf{x}_i$	Input feature vector for the $i$ -th observation
$y_i$	Target variable for the $i$ -th observation

$(\mathbf{x}_i, \mathbf{y}_i)$	$i$ -th observation consisting of input features and target
$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$	Dataset of $N$ observations
$\phi(\mathbf{z})$	Activation function applied to input $\mathbf{z}$
$\phi^{(i)}(\mathbf{x})$	Activation function at the $i$ -th layer of a deep neural network (DNN).
$g(\mathbf{x}; \boldsymbol{\theta})$	Neural network layer/module parameterised by $\boldsymbol{\theta}$
$g^{(i)}(\mathbf{x}; \boldsymbol{\theta})$	$i$ -th layer of a deep neural network, parameterised by $\boldsymbol{\theta}$
$\mathbf{h}^{(i)}$	Output at the $i$ -th layer of a deep neural network
$\mathcal{L}(\mathcal{D}; \boldsymbol{\theta})$	Loss/objective function for a model with parameters $\boldsymbol{\theta}$
$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}; \boldsymbol{\theta})$	Gradient of the loss function with respect to $\boldsymbol{\theta}$
$\eta$	Learning rate

## Chapter 1

# Introduction

The pursuit of human-like conversational agents has long captured the imagination of researchers, engineers, and the public alike. Long before the term *chatbot* was coined, philosophers and inventors speculated about mechanical companions capable of dialogue, from Yan Shi's humanoid described in ancient Chinese texts to Hephaestus' automata in Greek mythology. The fascination lies not merely in exchanging words but also in creating machines that appear intelligent, personable, and emotionally attuned. In 1950, Alan Turing reframed this centuries-old ambition as a scientific challenge: to build a programme that could sustain open-ended conversation indistinguishable from that of a human: passing a test later known as the "Turing Test" (Turing, 1950). Rather than settling philosophical debates on consciousness or self-awareness, it offered a practical criterion: the ability to impersonate a human in conversation convincingly.

The modern history of conversational agents traces back to programs such as ELIZA (Weizenbaum, 1966), developed in the 1960s, which mimicked the behaviour of a Rogerian psychotherapist using simple pattern matching and substitution. Although ELIZA demonstrated that surface-level language patterns could give an illusion of understanding, it had no model of the user's intent or emotions. In the 1970s, PARRY (Colby et al., 1972) extended this idea by simulating a patient with paranoid schizophrenia and adding handcrafted state variables to emulate a personality. Often described as the first system to pass a Turing-style test, PARRY underscored that beyond linguistic fluency, the modelling of personality and affective traits is central to creating more human-like interactions.

Early conversational programs such as ELIZA and PARRY relied heavily on manually engineered rules and state transitions. While these handcrafted approaches demonstrated that limited conversation could be simulated through predefined patterns, they also suggested the inherent complexity and ambiguity of human dialogue. The open-ended variability in human communication meant that the vision of an omnipotent, truly open-domain conversational agent remained far beyond reach.

In response to these challenges, both academic and industrial efforts shifted towards developing specialised dialogue systems designed for narrow, well-defined domains (Young et al., 2013). Such systems, often referred to as task-oriented dialogue (ToD) systems, focus on a constrained set of goals, enabling more effective optimisation for specific objectives such as factual accuracy, efficiency, and task success (Jurafsky and Martin, 2025). This focus on task success has clear practical advantages: it reduces the problem space, facilitates reproducible evaluation, and avoids the unpredictable complexity introduced by human-related factors such as emotional nuance. However, this pragmatic simplification also sidelines one of the original aspirations of dialogue research: the creation of agents that can interact with humans in an emotionally intelligent manner.

With advances in computation (Nickolls et al., 2008), deep learning (LeCun et al., 2015), and the availability of data (Budzianowski et al., 2018), conversational agents have become increasingly capable and versatile. The 2010s saw the proliferation of personal assistants such as Apple's Siri, Amazon's Alexa, and Google Assistant, capable of answering factual questions, executing commands, and managing simple tasks across multiple but still limited domains. More recently, the emergence of large language models (LLMs) such as ChatGPT (OpenAI, 2022) and Gemini (Team, 2024) has pushed the boundaries of instruction-following and open-domain conversation, showcasing an unprecedented breadth of language understanding and generation (Beeching et al., 2023). These advances open the door to revisiting the design of ToD systems with a renewed perspective: rather

than optimising solely for task success, we can begin to incorporate emotion awareness into their core capabilities.

In this thesis, we investigate how emotion can be integrated into ToD systems to make them more human-like, thereby enhancing user satisfaction from both functional and affective perspectives. By jointly addressing the informational and emotional aspects of interaction, we aim to bridge the gap between the efficiency of specialised systems and the richness of natural human conversation.

## 1.1 Task-oriented Dialogue Systems

ToD systems are designed to help users accomplish well-defined goals within specific domains, such as booking a hotel and finding a restaurant. Unlike open-domain conversational agents, which aim to handle any topic of conversation and primarily focus on engaging the user with natural and interesting responses, ToD systems aim to fulfil the user’s goal efficiently (Jurafsky and Martin, 2025).

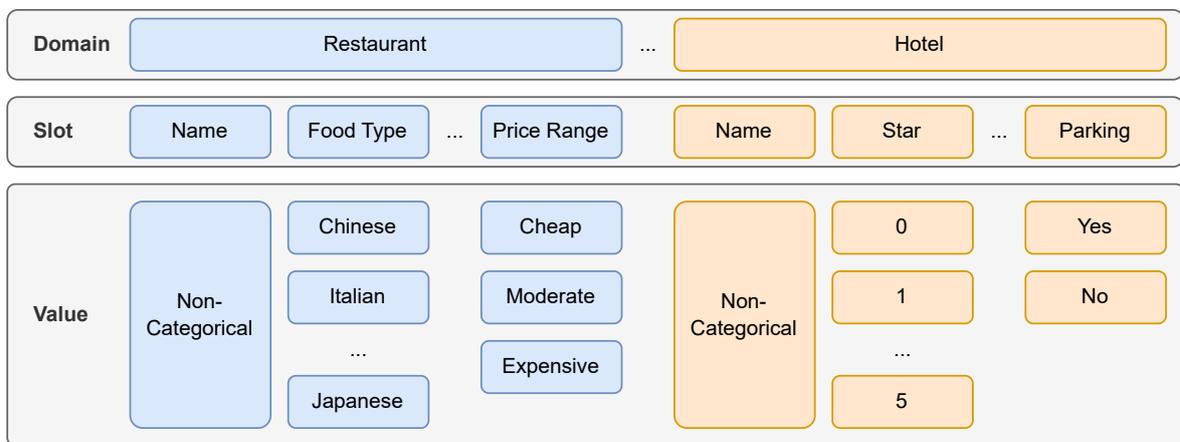


FIGURE 1.1: The ontology defines the scope of what a ToD system can understand and talk about.

Since the operation of ToD systems is centred around efficient task fulfilment, systems operate under a constrained ontology, which is hand-crafted by system builders to define the scope of information that can be understood and expressed by the system (Young, 2002). The ontology specifies the factual knowledge relevant to user goals and database entities. As illustrated in Figure 1.1, it includes domains (e.g., restaurants), domain-specific slots (e.g., restaurant price range), and slot-specific values (e.g., cheap, moderate, or expensive for the price range of restaurants).

Grounded in ontology, the system typically maintains a dialogue state, which estimates the user’s goal, tracks the progress of the goal completion, and represents search constraints. The underlying database contains entities represented as slot-value pairs. The dialogue state is used to query this database in a symbolic, rule-based manner to retrieve entities that satisfy the user’s requirements. Such a query process is illustrated in Figure 1.2.

Figure 1.3 shows a high-level overview of the user-system and system-database interactions. In a typical interaction, the user communicates their needs in natural language, which is inherently noisy, ambiguous, and variable. The system must interpret these utterances and maintain a dialogue state, which is a representation of the system’s belief of the user’s goal. The dialogue state, typically represented symbolically, supports precise, unambiguous interaction with an external knowledge base, which likewise stores information in a structured, symbolic format. The ultimate objective is to provide the user with one or more database entities (or take actions) that satisfy all of the user’s requirements, thereby completing the task successfully.

Traditionally, ToD systems have been implemented as modular pipelines, comprising the following main stages:

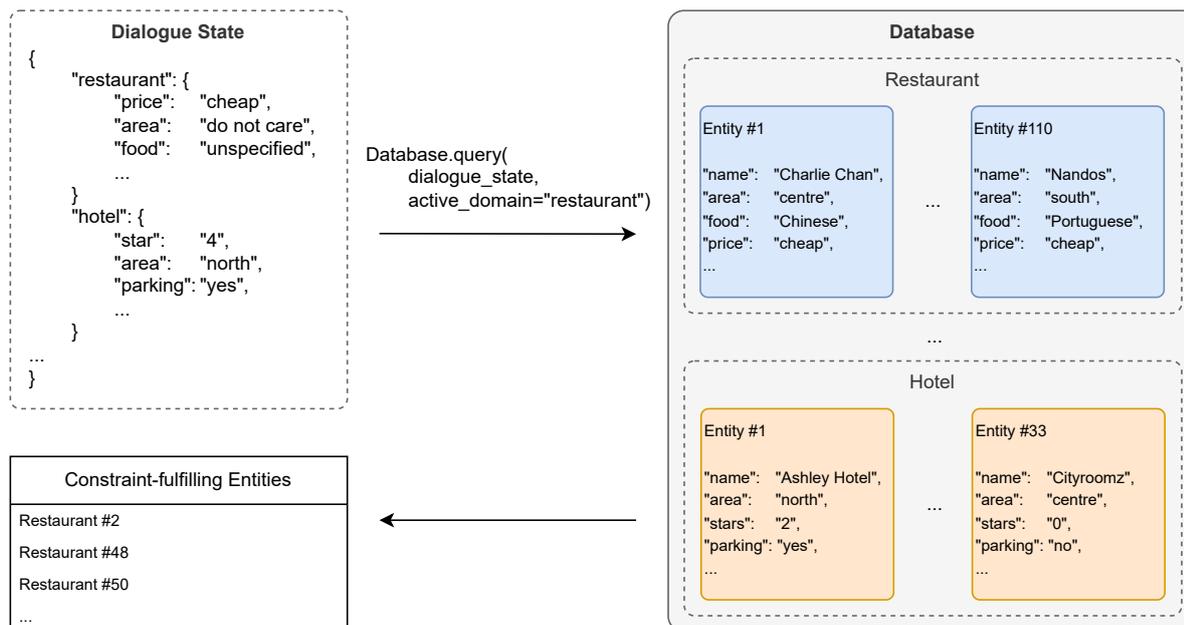


FIGURE 1.2: The ToD system constructs the dialogue state to capture the dialogue context and the user's goals and constraints. This state is subsequently used to form database queries that return entities fulfilling the user's search constraints.

- **Natural Language Understanding (NLU)** parses the user's utterance into a structured semantic representation, such as dialogue acts or slot-value pairs.
- **Dialogue State Tracking (DST)** maintains an explicit representation of the current state of the conversation, including accumulated constraints, user goals, and relevant context. In practice, the task of NLU is usually integrated into DST (Heck et al., 2022; Niekerk et al., 2021).
- **Dialogue Policy** determines the next system action based on the current dialogue state, often implemented using rule-based strategies (Paek and Pieraccini, 2008) or learned via reinforcement learning from interactions with real (Gašić et al., 2013) or simulated (Lin et al., 2022) users.
- **Natural Language Generation (NLG)** converts the semantic system action from the policy into a natural language response.

This architecture facilitates modular development and debugging, but error propagation between components can limit overall system performance.

Researchers have explored end-to-end systems (Hosseini-Asl et al., 2020; Wen et al., 2017), which map directly from dialogue history to system response without explicitly separating intermediate components. Such models can be trained in a supervised fashion using human-human dialogue corpora.

The emergence of general-purpose LLMs has given rise to agentic ToD systems, which rely on in-context prompting rather than task-specific training to handle complex dialogue contexts. By leveraging task descriptions, examples, and dialogue history, these models can adapt to a wide range of tasks in zero- or few-shot settings (Heck et al., 2023; Lin et al., 2025). However, their versatility often comes at the cost of robustness and consistency, and they typically lag behind task-optimised models in high-stakes or domain-specialised scenarios (Feng et al., 2025).

The evaluation of ToD systems is typically centred on task success and dialogue efficiency (e.g., number of turns to completion) (Lee et al., 2018; Nekkinda and Dušek, 2021; Zhu et al., 2023). Task success is defined in terms of whether the system retrieves an entity that meets all user-specified constraints or makes bookings according to user requirements.

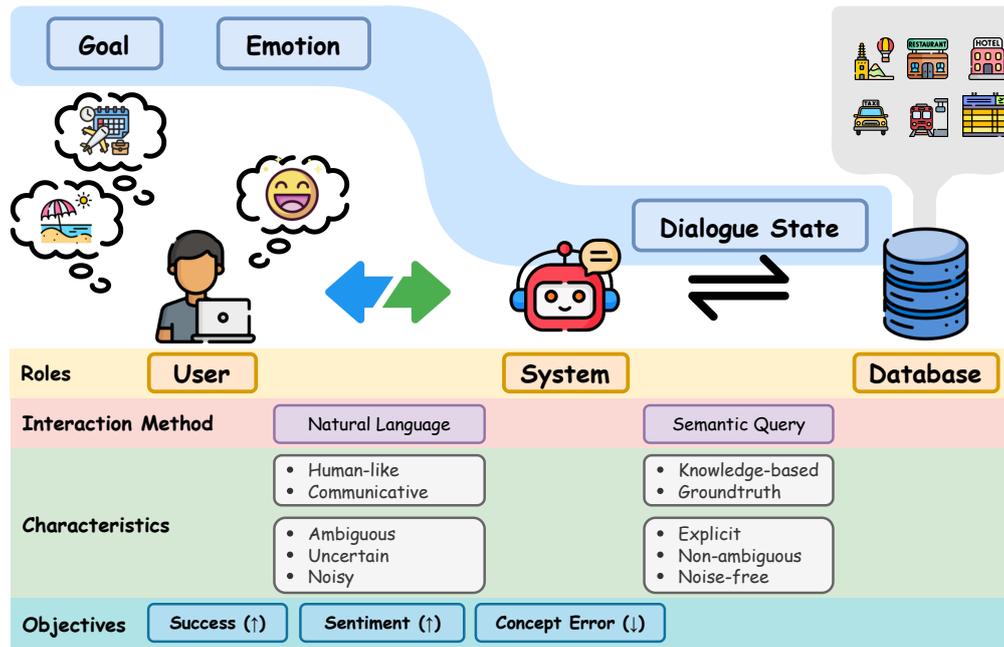


FIGURE 1.3: A ToD system interacts with the user in a noisy environment while querying the database symbolically. Its primary objective is task success, but it also aims to enhance the user’s emotional experience by accurately conveying factual knowledge with appropriate sentiment.

While these metrics offer clear and objective evaluation criteria, they primarily capture the functional dimension of the interaction. Prior evaluation frameworks have long recognised the importance of subjective user experience. For example, the PARADISE framework (Walker et al., 1997) evaluates the overall system performance by combining objective dialogue metrics with qualitative ratings from users reflecting their satisfaction. Similar approaches have been adopted in subsequent work, where user satisfaction or perceived quality is assessed through post-dialogue surveys or automatic assessment (Higashinaka et al., 2010). These are used either as an evaluation signal or as a training objective (Ultes et al., 2017).

However, in many such approaches, subjective experience is measured as a single, broad user satisfaction rating that aggregates multiple aspects of the interaction. These ratings often overlap with less subjective factors, such as automatic speech recognition errors or dialogue efficiency, and therefore provide limited insight into specific affective or emotional responses. Therefore, although user satisfaction is explicitly considered, the direct modelling of emotional dimensions of the user experience and the use of it as a system optimisation objective remain underexplored in standard ToD evaluation (Lin et al., 2023).

## 1.2 Emotionally Intelligent Task-oriented Dialogue Agent

Human conversation is inherently affective. Beyond the literal content of an utterance, interlocutors convey attitudes, emotions, and intentions that shape how their words are interpreted (Picard, 1997). Emotions influence the choice of words (Mohammad and Turney, 2010), prosody (Frick, 1985), and non-verbal cues (Busso et al., 2008), while also guiding turn-taking (Stevanovic and Peräkylä, 2015), emphasis (Banse and Scherer, 1996), and the interpretation of ambiguous or underspecified statements (Forgas, 1995). These affective signals serve important communicative functions: they help regulate social interaction (Ekman, 1992a; Parkinson, 1996), reach consensus or resolve disagreement (Dijk et al., 2008), and maintain interpersonal rapport (Acosta, 2009).

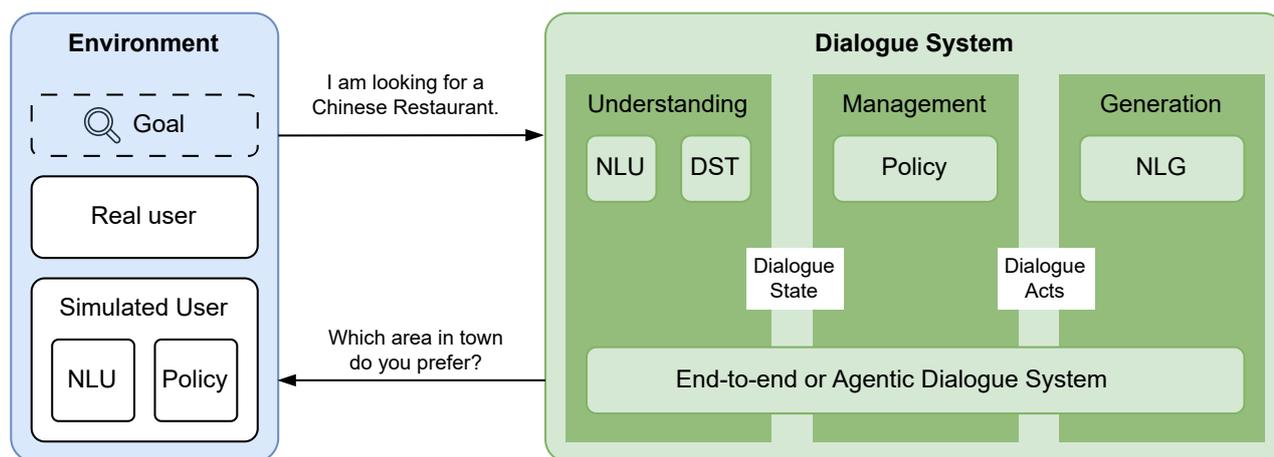


FIGURE 1.4: ToD systems follow either a modular or an end-to-end architecture, employing key modelling stages of understanding, management, and generation.

### Role of Emotion in Task-oriented Dialogues

For dialogue agents, this means that purely transactional exchanges, although sufficient to accomplish factual goals, may fall short in providing a satisfying and engaging user experience (Feng et al., 2024a). The ability to detect, interpret, and respond appropriately to the emotional states of users is not only desirable in open-domain or chit-chat systems (Zhou et al., 2018) but is also crucial in ToDs, where users' affective states can impact both their willingness to continue the interaction and the likelihood of successful task completion (Feng et al., 2022).

Although ToD systems have traditionally been evaluated and optimised for task success, the affective dimension of interaction can directly influence functional outcomes from the following perspectives:

- **Error recovery and frustration management:** A user who is repeatedly misunderstood may become frustrated and abandon the interaction unless the system detects and addresses their emotional state.
- **Recommendation:** In domains such as travel planning, users may be indecisive or hesitant. The system should recognise these cues and proactively offer empathetic, contextually appropriate recommendations to guide the user toward a decision.
- **User retention and satisfaction:** Positive emotional engagement can improve user satisfaction and promote continued use of the system.

For example, a restaurant-booking system could detect that a user is sounding frustrated after several failed search attempts and adapt by relaxing constraints, offering alternative suggestions, or providing a more empathetic acknowledgement.

### Key Abilities of Emotional Intelligence

The emotional intelligence of humans was first defined by Psychologists Peter Salovey and John Mayer as the ability to understand the meaning of emotions, to identify one's own and others' feelings, to use this information to guide thought and action, and to manage emotions for appropriate reactions (Salovey and Mayer, 1990). This was later conceptualised in the theory of affective computing by Rosalind Picard (1997) as three core abilities of an emotional computer: to perceive, to process, and to express emotions. These align to three stages of the traditional ToD pipeline as illustrated in Figure 1.4:

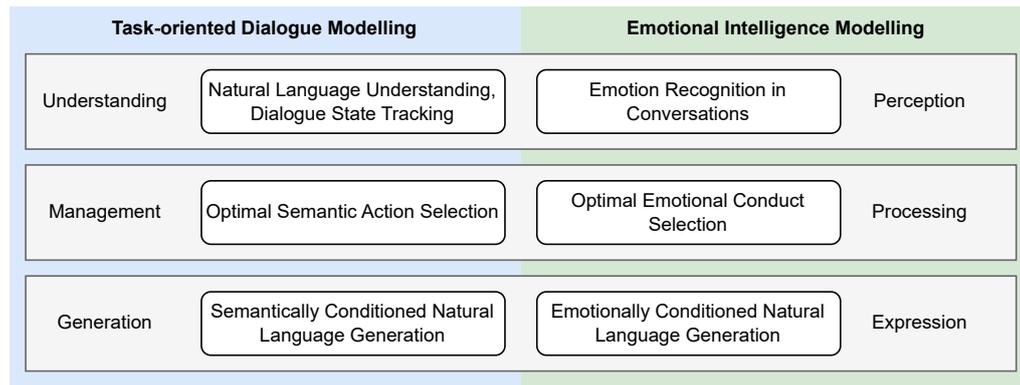


FIGURE 1.5: Capabilities of emotional intelligence can be mapped alongside the ToD modelling pipeline and formulated as computational tasks.

1. **Emotion Perception:** Analogous to NLU and DST, this involves recognising and interpreting the user’s affective state from their utterances and other contextual signals. Techniques range from generic sentiment analysis approaches (Pang et al., 2002) to dedicated emotional modelling in ToDs (Feng et al., 2023).
2. **Emotion Processing:** Comparable to dialogue action selection by the policy, this refers to deciding how the system should adapt its conversational strategy in light of the detected emotion and choosing the optimal emotion in the response.
3. **Emotion Expression:** Mirroring NLG, this entails producing responses that convey appropriate affective cues, such as compassion and enthusiasm, while remaining task-focused. This can involve lexical choice and emotional act selection (Zhou et al., 2018).

The analogy between ToD system modules and emotional intelligence capabilities, as outlined in Figure 1.5, suggests a principled approach to integrating emotion into ToD systems and motivates the research works introduced in this thesis.

### Multimodality of Emotion

Human emotion is inherently multimodal. In spoken or face-to-face interactions, emotion is often conveyed implicitly through tone, rhythm, facial expressions, and posture, sometimes even more strongly than through words themselves (Ekman, 1992b). As a result, emotional signals from multiple modalities can provide richer information for robust emotion perception, processing, and expression. There have been extensive studies in the field of affective computing (Picard, 1997).

However, ToD systems deployed in real-world applications, such as customer support chatbots, booking assistants, and information services, are predominantly text-based. Users may express frustration, uncertainty, or satisfaction indirectly through repetition, hedging, politeness strategies, or deviations from task-focused language, rather than through explicit emotion words or paralinguistic signals (Brown and Levinson, 1987). This makes emotion recognition and expression substantially more challenging.

This thesis deliberately focuses on text-based ToD to address this practical and underexplored challenge. By constraining the modality to text, we investigate how emotional information can be modelled, annotated, and utilised when only linguistic and contextual signals are available. This focus not only reflects the dominant deployment scenario of contemporary ToD systems, but also enables controlled analysis of the role of emotions in ToDs, without confounding multimodal cues.

While multimodal emotion modelling remains an important direction for future research, advancing emotionally intelligent behaviour in text-based dialogue systems is a necessary first step toward more applicable and emotionally intelligent ToD agents.

## 1.3 Challenges in Building Emotionally Intelligent Task-oriented Dialogue Systems

While the concept of emotionally intelligent ToD systems is compelling, its practical implementation faces several key challenges:

### 1.3.1 Data Scarcity and Annotation Challenges

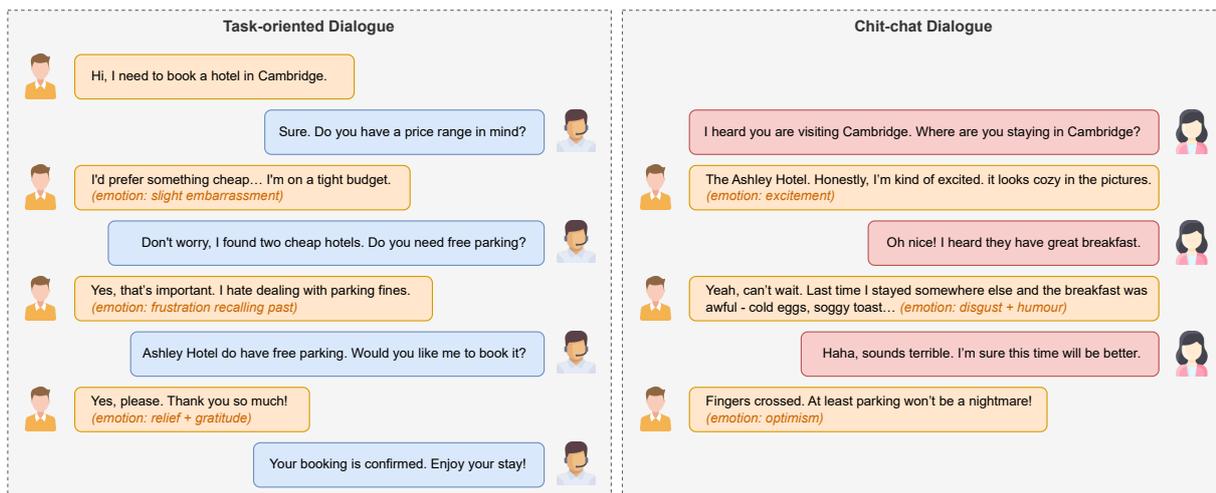


FIGURE 1.6: Illustration of a ToD (hotel booking) versus a chit-chat dialogue (talking about the booked hotel). In ToD, emotions are conveyed implicitly and tied to task progress (e.g., urgency, relief), whereas in chit-chat they are expressed more openly, reflecting personal experiences and social sharing.

The development of emotionally intelligent ToD systems is hindered by the lack of suitable data resources. Substantial research efforts have been invested into inherently emotion-rich conversation scenarios such as chit-chat, which serves as a medium for sharing personal experiences (Li et al., 2017), and psychological interviews, which explore a specific type of affective mental state (Gratch et al., 2014). By contrast, emotions in ToDs are less frequent, more subtle, and embedded within goal-driven exchanges. Figure 1.6 shows how emotions are conveyed in a ToD and a chit-chat dialogue, respectively, under the same topic. Hence, comparatively less research attention has been attracted to emotions in ToDs. In addition, emotions are inherently subjective, context-dependent, and are often more subtle in ToDs, making consistent annotation particularly challenging. Moreover, although large-scale datasets for emotion recognition exist in open-domain or chit-chat dialogues (Busso et al., 2008; Li et al., 2017), these resources do not transfer well to ToDs. In the task-oriented set-up, emotions are typically embedded in goal-directed interactions and arise indirectly rather than being the primary focus of the interaction. They are shaped by task progress, success, or system failures. As a result, generic emotion taxonomies such as Ekman's six basic emotions (Ekman, 1971) or dimensional models like the circumplex framework (Russell, 1980) fail to capture the nuanced affective states that emerge in ToD contexts. For example, user impatience due to system errors, relief after task completion, or uncertainty when goals are underspecified. Dedicated annotation schemes are therefore required to accurately represent these nuanced expressions of emotion in ToDs.

Overall, to advance research in this area, there is a pressing need for large-scale, high-quality corpora with carefully designed annotation schemes tailored to ToDs. Such resources would provide a foundation for training and evaluating emotionally intelligent systems that can recognise, interpret, and respond appropriately to user emotions during task completion.

### 1.3.2 Integration of Emotion into Task-oriented Dialogue Systems

As discussed in Section 1.2, integrating emotion into ToD systems requires the system to understand, process, and express emotions. Each step presents unique challenges.

#### Emotion Perception

Emotion understanding requires the system to track and predict users' affective states as they evolve over multiple dialogue turns. This necessitates an effective and robust emotion recognition in conversation (ERC) model. However, user emotions in ToDs differ from those in chit-chat dialogues. Specifically, user emotions arise from the goal-directed conversational context. They are subtler as the sharing of emotional experience is not the main purpose of the conversation, and they are centred around the task. The vocabulary in ToDs includes more functional words and named entities, differing vastly from chit-chat dialogues (Stricker and Paroubek, 2023). Existing ERC methods, typically optimised for open-domain or casual dialogues, may fail to capture these task-specific emotional cues or adapt to the linguistic styles in ToDs (Feng et al., 2022, 2023). Consequently, a specialised ERC model tailored to the task-oriented context is essential for enabling the system to select emotion-aware actions and providing accurate learning signals for dialogue policy optimisation.

#### Emotion Processing

In task-oriented interactions, users' emotions often provide critical signals about their satisfaction, frustration, or uncertainty, which can directly impact task completion. Like human operators in such scenarios, ToD systems can use these as feedback signals to guide their thinking and actions, aiming to achieve more efficient user goal achievement and to enhance user experience. Therefore, emotion processing should not be treated as an independent module but is most effectively integrated into the dialogue policy to determine the system's most appropriate emotional conduct alongside functional dialogue actions. This integration requires balancing task success with the user's emotional experience: emotion-aware adaptations must still ensure accurate and efficient task completion. Overemphasising emotional expressiveness at the expense of functional progress can reduce perceived competence by the user. Therefore, careful design of learning signals and optimisation objectives is necessary to simultaneously achieve both task effectiveness and emotional appropriateness.

#### Emotion Expression

In human communication, emotions are not only recognised and processed but also actively expressed through language, prosody, and style, shaping how messages are perceived. For ToD systems, the primary goal of delivering accurate information imposes an additional and strict requirement on emotion expression: emotion must be conveyed appropriately without altering the factual or semantic content of the intended response. For example, a negative sentiment in the system's reply is often associated with temporary task failure (e.g., no available options matching the search constraints or a booking failure), but it may also serve as an empathetic acknowledgement of the user's frustration. Achieving this balance requires disentangling affective information from task-related semantics during natural language generation so that the emotional tone can be adapted without distorting the informational content of the dialogue. At the same time, introducing affective variation increases the risk of hallucinations, where emotional framing may inadvertently change factual details or generate unsupported information (Xing et al., 2025). Thus, emotion-aware natural language generation must be carefully constrained to ensure that empathy enhances user experience without compromising the reliability and consistency of task-related outputs.

Together, emotion perception, processing, and expression form a tightly coupled pipeline in emotionally intelligent ToD systems. Errors or limitations in any component can propagate and compromise both task effectiveness and the user's emotional experience. Designing systems that

jointly optimise these elements remains a central research challenge: the system must accurately perceive subtle emotions, select context-appropriate responses, and express them naturally, all while reliably achieving the user’s task goals. Addressing this interplay is critical for building dialogue agents that are not only functionally competent but also socially and emotionally attuned.

## 1.4 Contributions

In this thesis, we address the core challenges discussed in Section 1.3 and advance the design of emotionally intelligent ToD systems through a sequence of works. The contributions progress in three stages, from constructing resources for modelling user emotion, to developing robust recognition methods, and finally to integrating affect into the full ToD pipeline.

### 1.4.1 Dataset Construction for Emotion Modelling in Task-oriented Dialogues

The foremost challenge in developing emotionally intelligent ToD systems is the absence of resources that capture affective behaviours specific to task-oriented interactions. Existing emotion corpora largely reflect chit-chat or psychological settings, while the few ToD-related datasets available focus on generic sentiment polarity and fail to capture nuanced states such as frustration from repeated booking failures, uncertainty when goals are underspecified, or relief upon task completion. To address this, we take inspiration from the emotion model of Ortony, Clore, and Collins (OCC, Ortony et al., 1988) and design a ToD-specific user emotion taxonomy and annotation protocol tailored to the dynamics of task progress. Building on this framework, we construct EmoWOZ, a large-scale corpus containing both human-human and human-machine dialogues, annotated with high-quality labels via a series of quality assurance procedures. EmoWOZ covers a wide behavioural spectrum of user emotions in ToD and enables benchmarking of existing ERC models. Our experiments reveal challenges unique to ToD emotion recognition. We also demonstrate that incorporating ERC as an auxiliary task can improve dialogue state tracking, highlighting its value beyond a stand-alone classification dataset. This contribution provides the first reproducible foundation for emotion recognition in ToDs, enabling systematic evaluation and supervised training of emotion-aware ToD modules. The research work has been presented in the following publication:

- Shutong Feng et al. (June 2022). “EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4096–4113. URL: <https://aclanthology.org/2022.lrec-1.436/>

### 1.4.2 Emotion Recognition Learning in Task-oriented Dialogues

Benchmarking on EmoWOZ highlights that recognising emotions in ToD remains challenging because they are often sparse, subtle, and tightly intertwined with task progression. Conventional methods developed for chit-chat data struggle to capture these nuances, leading to limited robustness and poor generalisation. We propose a framework for adapting chit-chat ERC models to the task-oriented domain by addressing three critical aspects of learning: data, features, and objective. To mitigate data imbalance, we employ data augmentation techniques to improve the recognition of rare emotions. To incorporate task-relevant information, we introduce dialogue states as auxiliary features. Further leveraging the annotation scheme of EmoWOZ and interconnections between labels, we devise a multi-task learning objective and a novel emotion-distance weighted loss function. Putting together these strategies significantly enhances the ERC performance. This contribution establishes practical methods for building accurate and robust ERC models in ToD, consolidating the foundations laid by EmoWOZ. The research work has been presented in the following publication:

- Shutong Feng et al. (Sept. 2023). “From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue”. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Svetlana Stoyanchev et al. Prague, Czechia: Association for Computational Linguistics, pp. 85–103. DOI: 10.18653/v1/2023.sigdial-1.8. URL: <https://aclanthology.org/2023.sigdial-1.8/>

While task-specific models are effective, they are limited by data availability and often lack cross-domain generalisation. The emergence of LLMs offers an opportunity to leverage broad linguistic and contextual knowledge for affect recognition. This naturally raises a key question: are dedicated data collection and task-specific adaptation still necessary for modelling emotions in conversation? To explore this, we investigate a range of open-source and proprietary LLMs as emotion recognisers under zero-shot, few-shot, and data-scarce conditions. We also explore affective behaviours beyond the scope of emotions in ToDs: emotions in chit-chat and mental state in psychological interviews. In addition, we evaluate robustness under automatic speech recognition (ASR) errors to simulate realistic noisy environments. Our findings show that while LLMs do not universally surpass specialised ERC models, they demonstrate notable robustness to ASR noise and adaptability in low-resource or unseen scenarios. Nevertheless, lightweight supervised models remain more suitable for high-stakes or time-sensitive applications. This work broadens the methodological toolkit for ToD ERC, highlighting LLMs’ complementary strengths in flexibility and robustness. The research work has been presented in the following publication:

- Shutong Feng et al. (Sept. 2024b). “Affect Recognition in Conversations Using Large Language Models”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, pp. 259–273. DOI: 10.18653/v1/2024.sigdial-1.23. URL: <https://aclanthology.org/2024.sigdial-1.23/>

### 1.4.3 Integrating Emotion into Task-oriented Dialogue Systems

Our work on dataset construction and emotion recognition learning lays the foundation for integrating emotions into ToD systems. We are the first to incorporate emotion into the full ToD pipeline, spanning natural language understanding, dialogue management, and response generation, optimised through an emotional user simulator. Specifically, we employ an ERC module alongside the DST module and extend the dialogue state to include the estimate of the user’s emotion. To handle the system’s affective behaviours, we complement EmoWOZ with system affective behaviour labels and expand the dialogue policy action space. The policy is then optimised with a multi-objective emotion-aware optimisation strategy that balances task success with affective appropriateness, trained via online reinforcement learning (RL) in natural language interactions. Subsequently, we condition the NLG both semantically and emotionally. We also develop supervisedly trained end-to-end counterparts. Experiments with both simulated and real users show that incorporating emotion improves task success and user emotional experience in both modular and end-to-end ToD systems. Notably, we also find that RL reduces hallucination in NLG, suggesting improved disentanglement between semantic content and emotional tone. This contribution demonstrates that ToD systems can be both effective and emotionally intelligent, offering practical design strategies for affect integration. The research work has been presented in the following publication:

- Shutong Feng et al. (Sept. 2024a). “Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, pp. 699–717. DOI: 10.18653/v1/2024.sigdial-1.60. URL: <https://aclanthology.org/2024.sigdial-1.60/>

While our prior work affirms the feasibility of emotion-aware ToD systems, further challenges remain in searching for the optimal choice of system architectures, representations of information, and

optimisation strategies. Therefore, we systematically explore a wide range of design considerations for emotionally intelligent ToD systems, including novel combinations not considered in prior research. Building on recent advances in natural language processing (NLP), we propose LUSTER, a unified framework that leverages LLMs as the backend, employs a fully lexicalised representation, and integrates online RL with multi-objective rewards that balance short-term emotion and long-term task success. This contribution offers a principled exploration of design choices and provides a holistic framework that unifies efficiency, robustness, and emotion in ToD systems. The research work has been presented in the following work, which is currently under review for publication:

- Shutong Feng et al. (2025). *Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation*. arXiv: 2507.01594 [cs.CL]. URL: <https://arxiv.org/abs/2507.01594>

## 1.5 Thesis Structure

This thesis consists of ten chapters:

- Chapter 1** presents a general introduction, setting the context and outlining the thesis objectives.
- Chapter 2** summarises techniques for data-driven machine learning, which are at the core of our approaches.
- Chapter 3** introduces theory and techniques for affective computing.
- Chapter 4** provides an overview of fundamental dialogue theories and their link to typical dialogue system set-ups.
- Chapter 5** introduces EmoWOZ, the first open-source and large-scale corpus with a dedicated labelling scheme for emotion recognition in ToDs.
- Chapter 6** introduces a general framework to adapt emotion recognition in conversation models to the task-oriented context.
- Chapter 7** investigates the use of large language models for recognising assorted types of affect under diverse dialogue settings, including ToDs.
- Chapter 8** introduces the incorporation of emotion into the complete ToD processing loop involving understanding, management, and generation.
- Chapter 9** investigates systematically into various design considerations of ToDs and proposes a unified architecture and multi-objective optimisation scheme for large-language-model-based end-to-end emotionally intelligent ToD systems.
- Chapter 10** summarises the thesis’s key findings, examines the implications of the presented research, and suggests directions for future investigation.



## Chapter 2

# Learning from Data

### 2.1 Overview

The field of natural language processing (NLP) has undergone a paradigm shift from rule-based systems toward data-driven machine learning approaches. Early NLP research was dominated by symbolic techniques, also known as Good Old-Fashioned Artificial Intelligence (GOFAI), where linguistic knowledge was encoded in handcrafted grammars, lexicons, and logical rules (Allen, 1995; Winograd, 1972). While these systems achieved partial success in constrained domains, their rigidity and poor scalability limited their applicability to real-world, open-domain tasks.

The rise of statistical learning in the 1990s marked the first major step toward data-driven NLP (Manning and Schütze, 1999). This shift was largely driven by the increasing availability of large text corpora, which allowed researchers to move beyond manually created rules. Instead of hardcoding every linguistic rule, statistical NLP leverages probability and machine learning to infer patterns and rules directly from the data. By treating language as a stochastic phenomenon, statistical models such as hidden Markov models and maximum entropy classifiers demonstrated strong performance in part-of-speech tagging, speech recognition, and information extraction. These methods highlighted the centrality of annotated corpora as the foundation of NLP progress.

The subsequent decade brought advances in neural networks and deep learning, which reconceptualised NLP as a problem of function approximation over large datasets (LeCun et al., 2015). Neural architectures, particularly recurrent neural networks (RNNs), and more recently, Transformers (Vaswani et al., 2017), enabled the automatic learning of distributed representations of language, eliminating the need for extensive manual feature engineering. This transformation has been especially impactful in dialogue systems, where data-driven methods capture nuances of intent, discourse structure, and affective state.

At the heart of these developments lies the principle of learning from data. Given a dataset of linguistic inputs and supervision in the form of labels or rewards, models learn mappings from text to representations or actions. Crucially, this learning paradigm encompasses a range of approaches - supervised, unsupervised, transfer, and reinforcement learning - that collectively form the methodological foundation for contemporary NLP research.

This chapter provides a systematic introduction to these paradigms. Section 2.2 discusses methods of data collection, the prerequisite for any learning-based system. Sections 2.3, 2.4 and 2.5 introduce neural networks as universal function approximators, optimisation techniques, and model architectures. Sections 2.6 to 2.10 detail specific learning paradigms: supervised learning, transfer learning, unsupervised learning and reinforcement learning. Section 2.11 gives special attention to large language models, which have revolutionised NLP in recent years. The thesis concludes in Section 2.12 with reflections on how these methods form the basis for emotionally intelligent task-oriented dialogue systems, the subject of this chapter.

## 2.2 Data Collection

The performance of data-driven NLP systems is fundamentally constrained by the availability and quality of data. Data collection, therefore, constitutes the first critical step in developing models for textual understanding and generation.

### 2.2.1 Source of Data

Data for NLP can be broadly categorised into naturally occurring corpora and constructed datasets. Naturally occurring corpora include web text (Common Crawl, Wikipedia), conversational logs, and social media data (Twitter, Reddit), which provide vast amounts of raw language. While these corpora generally lack explicit annotations, researchers often derive labels from their inherent structures. For example, Wikipedia’s infoboxes and knowledge graphs can be used to extract structured relationships for knowledge base construction (Auer et al., 2007). Social media content can also be analysed for sentiment or intent using built-in cues. For example, Felbo et al. (2017) leveraged emojis in Twitter to construct a sentiment analysis dataset, and Khodak et al. (2018) made use of a specific linguistic pattern, “/s”, to identify sarcasm in Reddit. In contrast, constructed datasets are curated with specific tasks in mind, serving as standardised benchmarks. Notable examples include Penn Treebank (Marcinkiewicz, 1994) for syntactic parsing or the Switchboard corpus (Godfrey and Holliman, 1993) for dialogue analysis.

A central trade-off in data-driven NLP lies between scale and quality. Large-scale datasets enable the training of foundation models with billions of parameters, but contain noisy, uncurated text. Conversely, high-quality expert-annotated corpora are typically small in scale. Hybrid approaches, where small annotated datasets are combined with large unlabelled corpora through task-specific fine-tuning or transfer learning, have proven highly effective in bridging this gap (Devlin et al., 2019).

In the context of dialogue systems, data often originates from human-machine (Levin et al., 2002) or human-human interactions (Godfrey and Holliman, 1993; Raux et al., 2005). The *Wizard-of-Oz method* is a popular approach for collecting realistic human-machine interaction data by having a human operator simulate the system’s responses (Budzianowski et al., 2018; Dahlbäck et al., 1993). Some datasets focus on a particular aspect of dialogue, such as the emotion of a speaker (Busso et al., 2008; Li et al., 2017). There are also datasets labelled for comprehensive modelling, such as the MultiWOZ dataset, which includes annotations for dialogue state, action, and goals in the task-oriented dialogue set-up (Budzianowski et al., 2018).

### 2.2.2 Annotation and Labelling

As we discuss later in Section 2.6, the supervised learning paradigm critically depends on annotated datasets, wherein raw linguistic inputs are paired with structured labels representing linguistic or semantic information such as intents, slots, dialogue acts, or affective states. The process of annotation may be conducted by domain experts, which typically yields highly reliable labels due to their specialised knowledge, but incurs substantial costs in terms of both time and resources. Alternatively, annotations can be obtained through crowdsourcing platforms (e.g., Amazon Mechanical Turk<sup>1</sup>, Prolific<sup>2</sup>), which enable large-scale data collection at a fraction of the cost. However, this approach often introduces challenges related to variability in annotator expertise, inconsistency in label assignment, and potential biases, thereby necessitating rigorous quality control mechanisms.

To obtain reliable labels, each data instance is usually annotated by multiple annotators. The quality of annotations is monitored through their agreement. Naïve measures such as raw percentage agreement  $\text{Agreement} = \frac{\# \text{Agreement}}{\# \text{Total Items}}$  are insufficient because they do not account for agreement by chance. In categorical tasks with skewed label distributions, annotators can appear to agree simply by favouring the majority class. For example, when labelling a dataset for sentiment analysis, annotators

<sup>1</sup><https://www.mturk.com/>

<sup>2</sup><https://www.prolific.com/>

may simply label each data sample as the most frequent emotion, which is usually neutral. This can create an inflated sense of reliability.

To evaluate the reliability of annotations and quantify the degree of agreement among annotators beyond chance, several inter-annotator agreement (IAA) metrics have been proposed. Three of the most widely used measures are Cohen's kappa, Fleiss's kappa, and Krippendorff's alpha. These metrics represent a progression in generality: from handling two annotators with nominal labels, to multiple annotators, and finally to heterogeneous settings with missing labels and different label types (nominal, ordinal, interval, ratio).

### Inter-annotator Agreement

**Cohen's Kappa ( $\kappa$ )** Cohen's kappa (Cohen, 1960) is designed for two annotators labelling the same set of items with categorical labels:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (2.1)$$

where the observed agreement  $p_o$  and expected agreement  $p_e$  are:

$$p_o = \frac{1}{N} \sum_{i=1}^C n_{ii}, \quad p_e = \sum_{i=1}^C \left( \frac{n_{i.}}{N} \cdot \frac{n_{.i}}{N} \right). \quad (2.2)$$

Here,  $N$  is the number of items,  $C$  is the number of categories,  $n_{ii}$  is the number of items jointly labelled as category  $i$ , and  $\{n_{i.}, n_{.i}\}$  are row and column marginals.

**Fleiss's Kappa ( $\kappa_F$ )** Fleiss's kappa (Fleiss, 1971) generalises Cohen's kappa to multiple annotators:

$$\kappa_F = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \quad (2.3)$$

where the mean observed agreement  $\bar{P}$ , the agreement for item  $i$   $P_i$ , and the expected agreement  $\bar{P}_e$  are

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i, \quad (2.4)$$

$$P_i = \frac{1}{n(n-1)} \sum_{c=1}^C n_{ic}(n_{ic} - 1), \quad (2.5)$$

$$\bar{P}_e = \sum_{c=1}^C p_c^2, \quad \text{with} \quad p_c = \frac{1}{Nn} \sum_{i=1}^N n_{ic}, \quad (2.6)$$

Here,  $N$  is the number of items,  $n$  is the number of annotators per item,  $n_{ic}$  is the number of annotators assigning item  $i$  to category  $c$ , and  $p_c$  is the overall proportion of category  $c$ .

**Krippendorff's Alpha ( $\alpha$ )** Krippendorff's alpha (Krippendorff, 1980) is the most general statistic. It handles multiple annotators, missing data, and different measurement levels (nominal, ordinal, interval, ratio):

$$\alpha = 1 - \frac{D_o}{D_e} \quad (2.7)$$

where

$$D_o = \frac{1}{N} \sum_{i=1}^N \frac{1}{n_i - 1} \sum_{k < l} \delta(c_{ik}, c_{il}), \quad (2.8)$$

$$D_e = \sum_{c_i} \sum_{c_j} p(c_i) p(c_j) \delta(c_i, c_j). \quad (2.9)$$

Here,  $n_i$  is the number of annotations for item  $i$ ,  $\{c_{ik}, c_{il}\}$  are the labels given by annotators  $k$  and  $l$ ,  $p(c)$  is the relative frequency of category  $c$ , and  $\delta(c_i, c_j)$  is a distance function (e.g.,  $\delta(c_i, c_j) = 1$  if  $c_i \neq c_j$ , else 0 for nominal data).

### Interpretation of Inter-annotator Agreement Metrics

All three inter-annotator agreement (IAA) metrics - Cohen's  $\kappa$ , Fleiss's  $\kappa_F$ , and Krippendorff's  $\alpha$  - share a similar interpretative framework. By construction, they correct for chance agreement, meaning that a value of 0 indicates that the observed agreement is no better than what would be expected by random labelling, while a value of 1 indicates perfect agreement. Negative values are possible and indicate systematic disagreement, where annotators agree less often than would be expected by chance.

For Cohen's  $\kappa$  and Fleiss's  $\kappa_F$ , the theoretical range is  $[-1, 1]$ . Krippendorff's  $\alpha$  has an upper bound of 1, but unlike  $\kappa$  it does not have a strict lower bound and can, in principle, take values below  $-1$  when disagreement is extreme. In practice, however, values typically fall between  $-1$  and 1.

A widely used set of benchmarks (Landis and Koch, 1977) interprets values as follows:  $< 0.00$  indicates poor or worse-than-chance agreement;  $0.00-0.20$  corresponds to slight agreement;  $0.21-0.40$  to fair agreement;  $0.41-0.60$  to moderate agreement;  $0.61-0.80$  to substantial agreement; and  $0.81-1.00$  to almost perfect agreement. Krippendorff (1980) further recommends that  $\alpha \geq 0.80$  be considered reliable for drawing conclusions, values in the range  $0.67 \leq \alpha < 0.80$  be treated as acceptable for tentative findings, and values below 0.67 be considered insufficient for scientific use.

### 2.2.3 Ethical and Practical Considerations

Data collection in NLP plays a foundational role in this thesis. It underpins our contributions, particularly the creation of EmoWOZ, the first large-scale dataset focusing on user emotions in task-oriented dialogue (Feng et al., 2022). At the same time, data collection raises several critical ethical concerns, including privacy, informed consent, and social bias, which we have carefully taken care of in all of our works.

**Privacy and informed consent.** The use of conversational data inherently intersects with issues of privacy and participant consent. Sensitive interactions, such as those in healthcare, mental health support, or customer service, often contain personal or confidential information (Sousa and Kern, 2023). Collecting and using such data without explicit consent risks violating individuals' rights to privacy, and even anonymised data may remain vulnerable to re-identification.

In our research, we explicitly mitigated these risks. The EmoWOZ dataset and its subsequently extended version (Feng et al., 2024a) were constructed from publicly available and ethically approved dialogue sources, with no inclusion of personally identifiable or sensitive user information. All human annotation tasks were reviewed and approved by the Ethics Committee of the Faculty of Mathematics and Natural Sciences of Heinrich Heine University Düsseldorf. Annotators were fully informed about the research purpose, data handling procedures, and potential risks (e.g., abusive language), and they provided explicit consent prior to participation. Moreover, all annotators' identities were fully anonymised. This ensures that no personal data or conversational content can be traced back to individuals.

**Bias and representativeness.** Datasets frequently encode social biases, including those related to gender, race, and cultural stereotypes, which are embedded in human language and social structures. For example, Bolukbasi et al. (2016) demonstrates how word embeddings can systematically reproduce gender stereotypes, while Blodgett et al. (2020) provides a comprehensive account of how bias in language technologies can perpetuate structural inequalities. Models trained on such data not only inherit these biases but also risk amplifying them when deployed at scale, potentially reinforcing stereotypes or inequalities.

In our data collection, we recruited annotators from the broad Amazon Mechanical Turk (AMT) worker pool, without restricting participation by gender, race, or cultural background. While this inclusive approach aimed to avoid narrowing the participant pool artificially, it does not guarantee the elimination of systematic biases, since AMT workers constitute a self-selected population with its own demographic skew. Nevertheless, based on our own experience in our data collection pre-trials, this strategy provided greater diversity than more restricted recruitment approaches, and was intended to help mitigate, to some extent, bias in emotion perception and labelling.

## 2.3 Neural Network Function Approximator

The success of modern NLP rests on the ability of neural networks to approximate complex, non-linear mappings between linguistic inputs and desired outputs. Unlike linear models, which are limited in expressive capacity, neural networks are capable of learning hierarchical and distributed representations of language. This property has positioned them as the dominant function approximators in contemporary NLP.

### 2.3.1 Neural Networks as Universal Function Approximators

The theoretical justification for neural networks largely stems from the universal approximation theorem (Cybenko, 1989; Hornik et al., 1989). This theorem asserts that a standard feedforward neural network, provided it has at least one hidden layer and employs a *non-linear activation function* such as the sigmoid, can approximate any continuous function defined on a compact subset of  $\mathbb{R}^n$  with arbitrary accuracy, given a sufficiently large number of hidden units. More formally, for any continuous function

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \quad (2.10)$$

and for any  $\epsilon > 0$ , there exists a neural network  $g_\theta$  such that

$$\sup_{x \in K} |f(x) - g_\theta(x)| < \epsilon, \quad (2.11)$$

where  $K \subset \mathbb{R}^n$  is compact, and  $\theta$  denotes the network parameters, i.e., the weights and biases.

It is important to note that the theorem guarantees *existence* rather than providing a constructive recipe for designing or training such a network. In other words, it tells us that some neural network configuration can approximate the desired function, but it does not specify how to efficiently find it or how many hidden units may be required in practice. Despite this limitation, the theorem provides a powerful theoretical foundation for viewing neural networks as general-purpose function approximators.

In the context of NLP, this principle has significant implications. Text-based tasks can often be framed as learning a mapping from sequences of discrete tokens to structured outputs. These outputs may take the form of class labels (e.g., emotion categories), continuous semantic representations (e.g., embeddings for similarity tasks), or even complex generative structures such as translated sentences or free-form text generation. The universal approximation property ensures, at least in theory, that neural architectures are capable of capturing these highly non-linear and compositional mappings, provided they are given sufficient capacity and appropriate training.

### 2.3.2 Non-linear Activation Functions

A crucial component of neural networks is the use of non-linear activation functions. Without non-linearities, a network composed of stacked transformations would collapse into a single linear mapping, thereby limiting its expressive power. The universal approximation theorem discussed earlier (Section 2.3.1) explicitly requires the presence of a non-linear activation function to guarantee the ability of neural networks to approximate arbitrary continuous functions.

Formally, consider a hidden unit activation:

$$\mathbf{h} = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}), \quad (2.12)$$

where  $\mathbf{x} \in \mathbb{R}^n$  is the input,  $\mathbf{W} \in \mathbb{R}^{m \times n}$  is the weight matrix,  $\mathbf{b} \in \mathbb{R}^m$  is the bias vector,  $\phi(\cdot)$  is a non-linear activation function applied element-wise, and  $\mathbf{h} \in \mathbb{R}^m$  is the resulting hidden activation vector. The choice of  $\phi$  significantly influences both representational capacity and trainability.

**Sigmoid and Hyperbolic Tangent (Tanh)** Early neural networks commonly employed the sigmoid function,

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (2.13)$$

which squashes real-valued inputs into the range  $(0, 1)$ .

Similarly, the hyperbolic tangent,

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad (2.14)$$

maps inputs into  $(-1, 1)$  and is zero-centered.

While useful for probabilistic interpretation, both functions saturate for large absolute values of  $x$ , making the training of deep networks difficult. We discuss the associated vanishing or exploding gradient problems in Section 2.5 when we talk about specific neural network architectures.

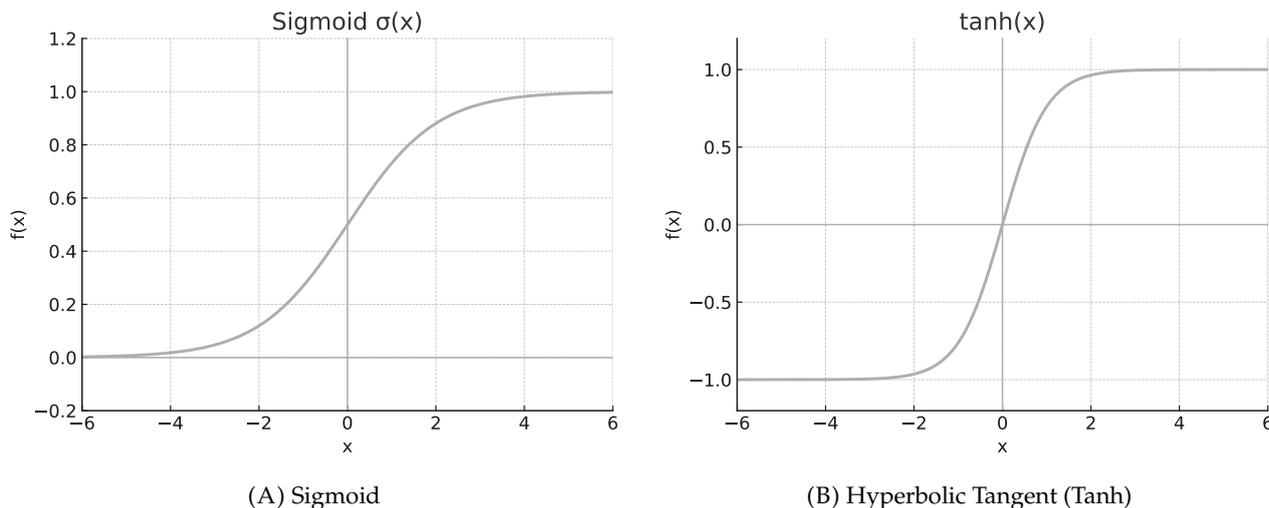


FIGURE 2.1: Illustrations of sigmoid and hyperbolic tangent activation functions.

**Rectified Linear Unit (ReLU)** The introduction of the Rectified Linear Unit (ReLU) addresses the vanishing gradient problem (Glorot et al., 2011). Defined as

$$\text{ReLU}(x) = \max(0, x), \quad (2.15)$$

it avoids saturation in the positive domain, mitigates vanishing gradients, and is computationally efficient. However, ReLU can lead to “dying neurons”, where units become inactive if their weights drive them into the negative domain persistently.

**Variants of ReLU** Several variants have been proposed to address these drawbacks. The Leaky ReLU (Maas et al., 2013) introduces a small slope  $\alpha$  in the negative domain:

$$\text{LeakyReLU}(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha x & \text{otherwise,} \end{cases} \quad (2.16)$$

where  $\alpha$  is a small constant (e.g., 0.1).

The Parametric ReLU generalises this idea by making  $\alpha$  learnable (He et al., 2015).

Another widely used alternative is the Gaussian Error Linear Unit (GELU, Hendrycks and Gimpel 2016). GELU is defined as

$$\text{GELU}(x) = x \cdot \Phi(x), \quad (2.17)$$

where  $\Phi(x)$  is the cumulative distribution function of the standard normal distribution. GELU has been particularly successful in Transformer-based architectures (Devlin et al., 2019; Radford et al., 2018).

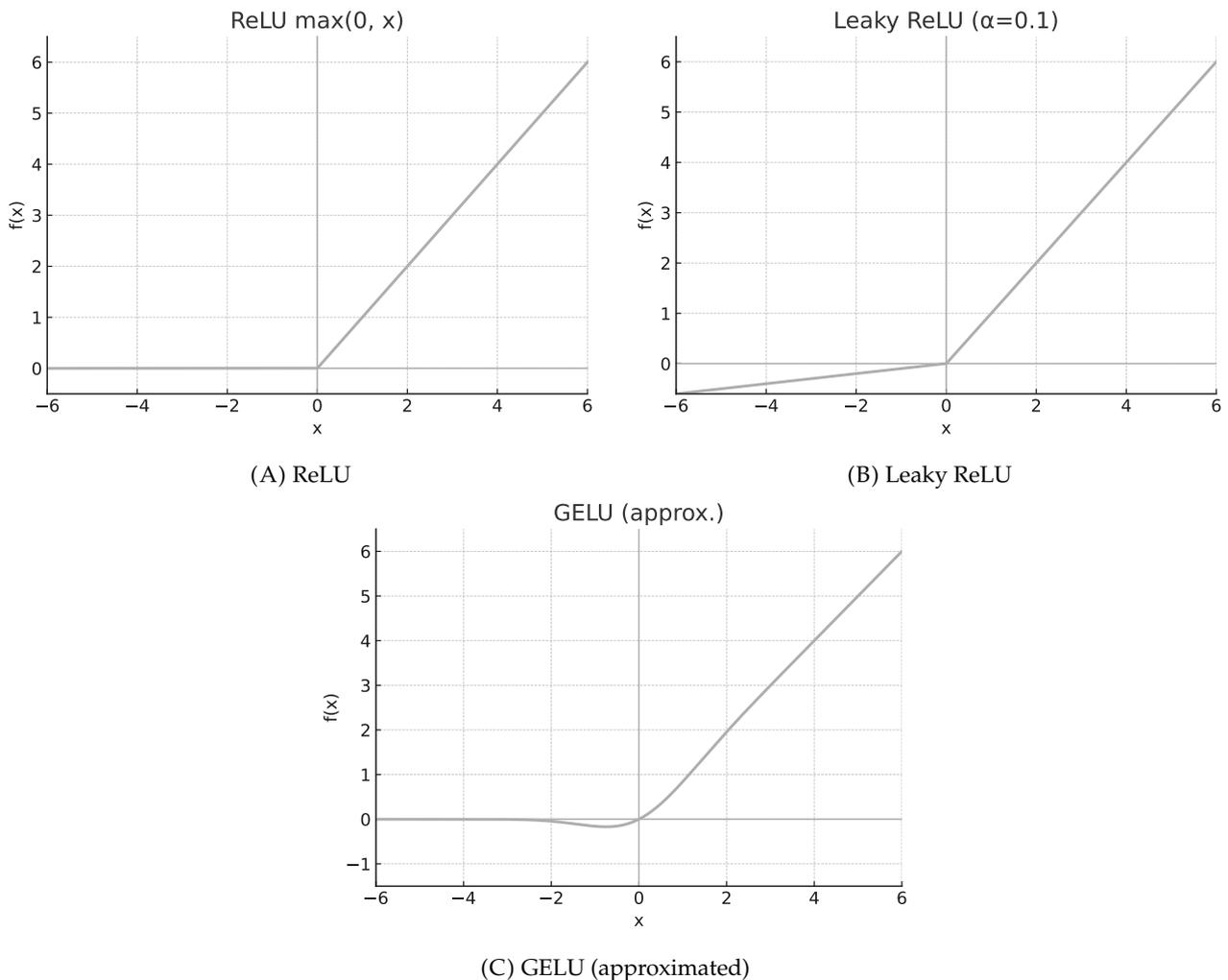


FIGURE 2.2: Illustrations of the ReLU activation function and its variants.

### 2.3.3 Representation of Text in Neural Models

Raw linguistic input is inherently symbolic and discrete, consisting of words or subword units drawn from a vocabulary. Neural models, however, operate in continuous vector spaces, where gradient-based optimisation techniques can be applied effectively. Bridging this gap requires a transformation from symbolic sequences to numerical vectors, typically achieved through embeddings. Formally, let a tokenised utterance be represented as

$$\mathbf{x} = (w_1, w_2, \dots, w_T), \quad w_t \in V \quad (2.18)$$

where  $T$  is the sequence length and  $V$  denotes the vocabulary. Each token  $w_t$  is then mapped to a dense vector representation in  $\mathbb{R}^d$ :

$$\mathbf{X} = (e_{w_1}, e_{w_2}, \dots, e_{w_T}), \quad e_{w_t} \in \mathbb{R}^d \quad (2.19)$$

where  $d$  is the embedding dimension and  $e_{w_t}$  is the embedding of token  $w_t$ .

Before the development of word embeddings, earlier NLP systems relied on more primitive text representations. The most common methods included one-hot encoding, where each word in the vocabulary is represented by a sparse binary vector of dimension  $|V|$  with a single one at the index corresponding to the word, and bag-of-words (BoW) or n-gram models, where text is represented by frequency counts of words or short sequences of words. While these approaches were simple and interpretable, they suffered from extremely high dimensionality, sparsity, and an inability to capture semantic similarity between words. For example, “dog” and “cat” would be orthogonal vectors in a one-hot space despite their semantic similarity.

The introduction of *distributed word representations* addressed these shortcomings. Word embeddings, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), represent words as dense, low-dimensional vectors learned from large corpora, where semantic and syntactic relationships are encoded geometrically. For instance, vector operations can capture analogies such as:

$$e_{\text{king}} - e_{\text{man}} + e_{\text{woman}} \approx e_{\text{queen}}. \quad (2.20)$$

These static embeddings assign a single vector to each word type, regardless of context. However, natural language meaning is inherently context-dependent. For example, the word “bank” in “river bank” and “financial bank” has completely different meanings with different preceding words and requires different representations. To address this limitation, modern approaches introduced contextual embeddings, where representations vary with surrounding text. Breakthroughs such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) demonstrated the effectiveness of deep, pre-trained language models that produce contextualised embeddings. These models leverage large-scale unsupervised training objectives, which we will discuss in Section 2.8, and have become the backbone of state-of-the-art natural language processing systems.

Text representation in neural models has evolved from sparse symbolic encodings to dense, distributional embeddings, and further to contextualised embeddings derived from pre-trained language models. This progression reflects the increasing ability of neural methods to capture the semantic and syntactic nuances of natural language.

## 2.4 Neural Network Optimisation

Section 2.3 introduces the building blocks of neural networks. In this section, we examine how to find the neural network that fulfils Equation 2.11. Optimising a neural network involves finding parameter values  $\theta$  that minimise a task-specific loss function  $\mathcal{L}(\theta)$ . The optimisation problem is generally formulated as

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\theta). \quad (2.21)$$

### 2.4.1 Loss Functions

Loss functions specify *what* a model should optimise during training by quantifying the discrepancy between predicted outputs and ground-truth labels. The choice of loss function depends on the task. Here are several fundamental loss functions that are widely adopted in NLP tasks.

#### Cross-entropy Loss

For supervised classification problems, the *cross-entropy (CE) loss* is widely adopted. Given a dataset of  $N$  training examples  $\{(x_i, y_i)\}_{i=1}^N$ , where  $y_i \in Y$  is the true label from the set of possible labels  $Y$ , the cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{|Y|} \mathbb{1}[y_i = c] \log p_{\theta}(y = c \mid x_i), \quad (2.22)$$

where  $\mathbb{1}[y_i = c]$  is an indicator function that equals 1 if the label is  $c$ , and  $p_{\theta}(y = c \mid x_i)$  is the predicted probability assigned by the model to class  $c$ . In practice, the probability is modelled with a softmax activation function in the output layer of the neural network, which we will discuss later in Section 2.6.2. This formulation encourages the model to assign high probability to the correct class and low probability to incorrect classes.

#### Mean Squared Error Loss

For regression tasks such as in the work of He and Lin (2016) where a continuous numerical value is to be predicted, the *mean squared error (MSE) loss* is used. Mathematically, it is defined as:

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{N} \sum_{i=1}^N \|y_i - f_{\theta}(x_i)\|^2, \quad (2.23)$$

where  $y_i$  is the ground-truth continuous target and  $f_{\theta}(x_i)$  is the model prediction. This loss penalises the squared deviation between predictions and targets, encouraging the model to approximate continuous values accurately.

#### Custom Losses

Custom loss functions are frequently introduced in NLP to better reflect the nuances of particular tasks.

In multi-task learning, for instance, the losses of multiple tasks are often combined into a weighted objective so that the model learns shared representations while balancing potentially competing goals (Caruana, 1997). We will discuss the details of multitask learning in the context of transfer learning later in Section 2.7.3.

Other losses consider label similarity, where the penalty for misclassification is modulated by the semantic proximity of labels. For example, predicting “joy” instead of “happiness” should be penalised less than predicting “anger”.

We use custom loss functions in our work described in Chapter 6 and illustrate how loss functions can be tailored to encode domain knowledge, reflect task-specific requirements, and ultimately improve performance beyond what standard objectives can offer.

### 2.4.2 Optimisers

Having defined the loss functions that guide learning, the next step is to consider *how* these objectives are minimised in practice. In neural networks, this is typically achieved through gradient-based

optimisation, with stochastic gradient descent and its variants forming the backbone of modern training strategies.

### Gradient Descent and Stochastic Gradient Descent

Consider a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  for which we wish to find a point  $x^*$  that minimises its value. To reason about how to move towards such a minimum, we approximate  $f$  locally around an arbitrary point  $x \in \mathbb{R}^d$ . By Taylor's theorem presented in the book of Marsden and Tromba (2003), a first-order expansion yields

$$f(x + \Delta x) \approx f(x) + \Delta x^\top \nabla_x f(x), \quad (2.24)$$

where  $\Delta x \in \mathbb{R}^d$  is assumed to be small.

To ensure that  $f(x + \Delta x)$  does not exceed  $f(x)$ , one natural choice is to set the update step proportional to the negative gradient:

$$\Delta x = -\eta \nabla_x f(x), \quad (2.25)$$

with  $\eta > 0$  as the step size or learning rate. This choice decreases the function value since the gradient always points in the direction of steepest increase.

Applying this idea iteratively defines the update rule

$$x_{t+1} = x_t - \eta \nabla_x f(x_t), \quad (2.26)$$

starting from an initial guess  $x_0$ . This leads to the most basic optimisation algorithm for minimising the loss of a neural network parameterised with  $\theta$ : *gradient descent*. Parameters are updated iteratively (from step  $t$  to  $t + 1$ ) in the direction opposite to the gradient of the loss function with respect to the parameters:

$$\theta_{t+1} = \theta_t - \eta g_t, \quad (2.27)$$

where  $\eta$  is the learning rate controlling the step size, and  $g_t = \nabla_\theta \mathcal{L}(\theta_t)$  denotes the gradient of the loss with respect to  $\theta$ .

While *gradient descent* computes gradients using the entire dataset, this is computationally expensive for large corpora. Instead, *stochastic gradient descent (SGD)* estimates the gradient on a small mini-batch of examples, introducing stochasticity that often helps escape local minima and saddle points (Bottou, 2010).

To further accelerate convergence, SGD is often combined with *momentum* (Rumelhart et al., 1986b), which accumulates a moving average of past gradients:

$$m_t = \beta m_{t-1} + g_t, \quad (2.28)$$

$$\theta_{t+1} = \theta_t - \eta m_t, \quad (2.29)$$

where  $\beta$  is the momentum coefficient. It follows that when the current gradient  $g_t$  and the previous momentum term  $m_{t-1}$  share the same sign, for example, when small gradients accumulate in a consistent direction, the resulting update is amplified. In contrast, if their signs differ, as is typically the case during oscillations, the update is reduced.

### Adam and AdamW

A particularly influential optimiser is **Adam** (Adaptive Moment Estimation, Kingma and Ba 2017), which combines the ideas of momentum and RMSProp by maintaining exponentially decaying

estimates of both first-order and second-order moments of the gradients:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t, \quad (2.30)$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t \odot \mathbf{g}_t, \quad (2.31)$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1 - \beta_1^t}, \quad \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1 - \beta_2^t}, \quad (2.32)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t + \epsilon}}, \quad (2.33)$$

where  $\{\mathbf{m}_t, \mathbf{v}_t\}$  are estimates of the first and second moments of the gradients, and  $\{\beta_1, \beta_2\}$  are hyperparameters controlling exponential decay rates. Adam has become the de facto optimiser for deep learning due to its robustness and ability to handle sparse gradients.

However, the original Adam couples weight decay with the adaptive learning rate, which can lead to suboptimal regularisation. **AdamW** (Loshchilov and Hutter, 2019) addresses this by decoupling weight decay from the update rule, improving generalisation in large-scale models (Devlin et al., 2019; Radford et al., 2018; Raffel et al., 2020). Today, AdamW with carefully tuned learning rate schedules is the standard choice for fine-tuning large pre-trained language models.

### 2.4.3 Practical Optimisation Techniques

Beyond optimisers and loss functions at the core of neural network optimisation, several additional techniques have been developed to stabilise training, accelerate convergence, and improve the generalisation of deep neural networks. These techniques are crucial for scaling models to large datasets and deeper architectures, as commonly found in modern dialogue systems.

#### Normalisation

Normalisation methods mitigate internal covariate shift by rescaling intermediate activations. The most widely used approach is **batch normalisation** (Ioffe and Szegedy, 2015), which normalises each feature across a mini-batch:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad \mathbf{y}_i = \gamma \hat{\mathbf{x}}_i + \beta, \quad (2.34)$$

where  $\mu_B$  and  $\sigma_B^2$  are the batch mean and variance, and  $\gamma, \beta$  are learnable scale and shift parameters. Batch normalisation improves gradient flow, allows for higher learning rates, and reduces sensitivity to initialisation.

For sequential models introduced later in Section 2.5, **layer normalisation** (Ba et al., 2016) is usually preferred, as it normalises across features within a layer for each sample individually:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_x - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}}, \quad (2.35)$$

where  $\mu_L$  and  $\sigma_L^2$  are computed across hidden dimensions rather than across the batch.

#### Gradient Clipping

For recurrent and very deep models, exploding gradients can destabilise optimisation. This challenge is not limited to supervised learning: it is also prevalent in reinforcement learning, where high-variance gradient estimates, especially in policy gradient or actor-critic methods, can lead to unstable or divergent updates. A common countermeasure is *gradient clipping* (Pascanu et al., 2013), where

gradients are rescaled when their norm exceeds a threshold  $\tau$ :

$$g \leftarrow g \cdot \frac{\tau}{\max(\|g\|, \tau)}. \quad (2.36)$$

This technique prevents unstable updates while preserving gradient direction.

### Regularisation

Effective training also requires strategies to prevent overfitting and improve generalisation. Common approaches include  $\ell_2$  weight decay, dropout, and early stopping based on validation performance.

**Weight decay** (also known as  $\ell_2$  regularisation) adds a penalty proportional to the squared magnitude of the parameters:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2, \quad (2.37)$$

where  $\lambda$  is a hyperparameter controlling the strength of regularisation.

**Dropout** randomly sets a fraction  $p$  of activations to zero during training:

$$h_i^{\text{drop}} = \frac{m_i h_i}{1 - p}, \quad m_i \sim \text{Bernoulli}(1 - p), \quad (2.38)$$

where  $h_i$  is the activation of unit  $i$ , and  $m_i$  is a random mask. The scaling factor  $\frac{1}{1-p}$  ensures that expected activations remain the same at test time.

These methods ensure that the trained parameters  $\boldsymbol{\theta}$  not only minimise the training loss but also generalise well to unseen data.

## 2.5 Neural Network Architectures

Several neural architectures have been proposed for modelling linguistic data, each introducing progressively more sophisticated mechanisms for capturing dependencies in text. Below, we discuss the key models, their theoretical underpinnings, and the practical techniques that enabled their success in natural language processing.

### 2.5.1 Feed-forward Neural Networks

The feed-forward neural network (FNN) defines a function  $\hat{y} = f(x; \boldsymbol{\theta})$  that maps the input  $x$  to the output  $\hat{y}$  with trainable parameters  $\boldsymbol{\theta}$ . In practice, the FNN is implemented as a series of  $L$  fully connected (FC) layers  $f_1, f_2, \dots, f_L$  as illustrated in Figure 2.3.

$$f_l(\mathbf{h}_{l-1}) = \phi_l(\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l) \quad (2.39)$$

Each FC layer takes input  $\mathbf{h}^{(l-1)} \in \mathbb{R}^{d_{l-1}}$ , a real-number vector with dimension  $d_{l-1} \in \mathbb{N}$  from the output of the previous layer  $f_{l-1}$  and outputs  $\mathbf{h}^{(l)} \in \mathbb{R}^{d_l}$ . The FNN's input is  $\mathbf{h}^{(0)} = x$  and output is  $\mathbf{h}^{(L)} = \hat{y}$ . The FC layer  $f_l$  is parameterised with  $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$  and  $\mathbf{b}_l \in \mathbb{R}^{d_l}$  with an activation function  $\phi_l$  as discussed in Section 2.3.2.

Early neural NLP systems relied on FNNs (Collobert et al., 2011) where an input sentence fragment of fixed size (e.g. a context window of length  $2k + 1$ ) is first represented as concatenated word embeddings  $e_w$ :

$$\mathbf{x} = [e_{w_{t-k}}, \dots, e_{w_t}, \dots, e_{w_{t+k}}] \in \mathbb{R}^{(2k+1)d}, \quad (2.40)$$

A major limitation of FNNs in NLP is their reliance on a fixed input length. Since these models operate over a concatenated window of token embeddings, they are inherently restricted to capturing dependencies within that predefined span. As a result, they are unable to model linguistic phenomena that require reasoning over longer sequences, such as coreference or discourse-level coherence.

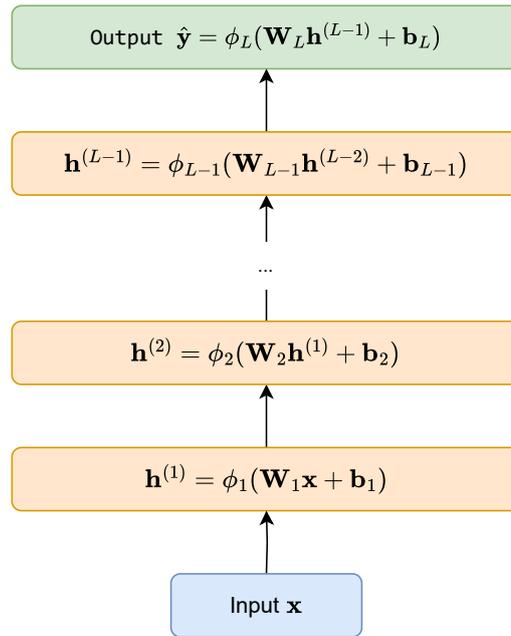


FIGURE 2.3: An illustration of a FNN architecture with input  $x$  and output  $\hat{y}$ . Each layer  $l$  is parameterised with a trainable weight matrix  $\mathbf{W}_l$  and a bias vector  $\mathbf{b}_l$ , and has a non-linear activation function  $\phi_l$ .

Another drawback is the absence of an explicit mechanism for temporal dynamics. Unlike recurrent or attention-based models, which we will discuss later, FNNs treat their input tokens as a flat vector without regard to order beyond the local window. This makes them poorly suited for tasks where sequential information is essential, such as language modelling or machine translation.

Nevertheless, despite these limitations, FNNs could successfully perform fundamental NLP tasks such as part-of-speech tagging and named entity recognition (ibid.). Crucially, they achieved this without the need for handcrafted linguistic features, thereby highlighting the potential of data-driven neural approaches as a departure from traditional feature-engineered systems.

### Gradient Vanishing and Exploding

While shallow FNNs can be trained effectively, very deep architectures, while benefiting from the increased number of parameters and therefore greater potential capability, suffer from the well-known *vanishing or exploding gradient problem*. During backpropagation, the gradient at a given layer is obtained by multiplying the gradients of subsequent layers. Formally,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{h}^{(l)}} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}^{(L)}} \prod_{j=l+1}^L \frac{\partial \mathbf{h}^{(j)}}{\partial \mathbf{h}^{(j-1)}}, \quad (2.41)$$

where  $\mathcal{L}$  denotes the loss,  $L$  the index of the final layer, and  $h^{(l)}$  the hidden representation at layer  $l$ .

This repeated multiplication of Jacobians can lead to gradients that shrink exponentially with depth (if eigenvalues  $< 1$ ) or grow uncontrollably (if eigenvalues  $> 1$ ). In practice, the shrinking case is more common due to saturating activation functions such as the sigmoid or tanh, resulting in earlier layers learning very slowly or not at all.

The *vanishing or exploding gradient problem* significantly limited the effective depth of early FNNs, making them difficult to scale to more complex tasks. Activation functions, initialisation schemes, layer normalisation, and residual connections were introduced to alleviate this issue, enabling the successful training of modern deep architectures.

## 2.5.2 Residual Neural Networks

One limitation of standard FNNs is the degradation of performance as the network depth increases. Although deeper models have greater capacity, they are more difficult to optimise due to vanishing gradients and optimisation instability. To address this, residual neural networks (ResNets) make it possible to train very deep architectures by introducing *skip connections* that allow gradients to flow more effectively during backpropagation (He et al., 2016) as depicted in Figure 2.4.

Formally, a residual block can be expressed as

$$\mathbf{h} = f_{\text{resnet}}(\mathbf{x}; \mathbf{W}, \mathbf{b}) + \mathbf{x}, \quad (2.42)$$

where  $\mathbf{x}$  is the input to the block,  $f_{\text{resnet}}(\mathbf{x}; \mathbf{W}, \mathbf{b})$  is a non-linear transformation parameterised by weights  $\mathbf{W}$  and biases  $\mathbf{b}$ , and  $\mathbf{h}$  is the block output. The term  $+\mathbf{x}$  denotes the skip connection, which bypasses the non-linear transformation and directly adds the input to the output.

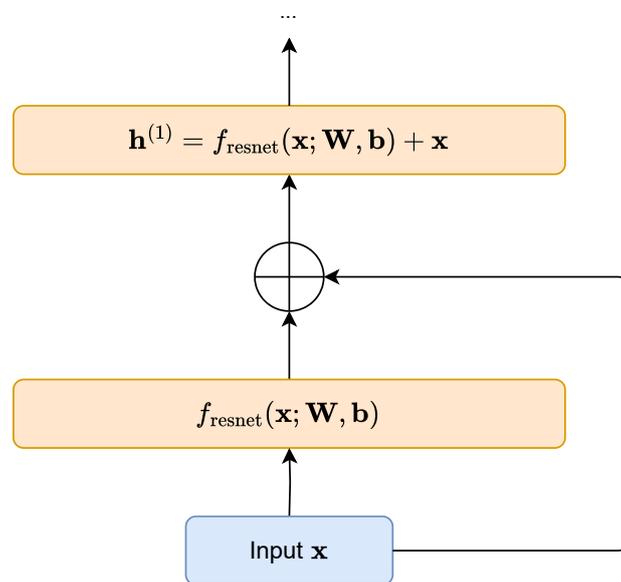


FIGURE 2.4: An illustration of a residual connection. The input  $\mathbf{x}$  skips and adds to the output of the non-linear residual block  $f_{\text{resnet}}(\mathbf{x}; \mathbf{W}, \mathbf{b})$  to output the hidden activation  $\mathbf{h}^{(1)}$ .

During backpropagation, the derivative through a residual block takes the form

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}} \left( \frac{\partial f_{\text{resnet}}(\mathbf{x}; \mathbf{W}, \mathbf{b})}{\partial \mathbf{x}} + \mathbf{I} \right), \quad (2.43)$$

where  $\mathbf{I}$  is the identity matrix. The identity term ensures that gradients can propagate directly through the skip connection, reducing the likelihood of vanishing gradients and allowing optimisation of much deeper networks.

Although originally proposed for computer vision, residual connections have been widely adopted in NLP architectures, especially Transformers, which we introduce in Section 2.5.4.

## 2.5.3 Recurrent Neural Networks

Neural network architectures discussed earlier, while useful, are inherently limited in their ability to capture sequential information. This is especially relevant to NLP as sentences and dialogues are inherently sequential. To overcome this, RNNs were proposed as a natural extension for modelling sequential data (Rumelhart et al., 1986a). The key idea behind RNNs is the use of a hidden state that

evolves over time, integrating information from both the current input and the history of past inputs. This hidden state acts as a form of memory, allowing the network to capture dependencies across time steps (Figure 2.5).

Formally, an RNN updates its hidden representation at time step  $t$  as

$$\mathbf{h}_t = \sigma(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{W}_x \mathbf{w}_t + \mathbf{b}_h) \in \mathbb{R}^m \quad (2.44)$$

where  $\mathbf{x}_t$  is the embedding of the  $t$ -th token,  $\mathbf{h}_t$  is the hidden state of dimension  $m$ , and  $\{\mathbf{W}_h, \mathbf{W}_x, \mathbf{b}_h\}$  are learnable parameters.

The output distribution over possible labels or tokens is then computed as

$$\hat{\mathbf{y}}_t = \text{softmax}(\mathbf{W}_y \mathbf{h}_t + \mathbf{b}_y) \quad (2.45)$$

where  $\{\mathbf{W}_y, \mathbf{b}_y\}$  map the hidden state into the output vocabulary or label space. The softmax function ensures that the output vector  $\mathbf{y}_t$  forms a valid probability distribution by exponentiating each logit and normalising across all classes:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^V e^{z_j}}, \quad (2.46)$$

where  $z_i$  denotes the logit corresponding to the  $i$ -th class, and the denominator sums over all  $V$  vocabulary items (or labels). Thus, each item  $\hat{y}_{t,i}$  in the output vector  $\hat{\mathbf{y}}_t$  is the normalised probability of predicting class  $i$  at time  $t$ , with the properties  $0 \leq \hat{y}_{t,i} \leq 1$  and  $\sum_{i=1}^{|V|} \hat{y}_{t,i} = 1$ . This allows the model to interpret  $\hat{y}_{t,i}$  as the probability of the  $i$ -th token or label.

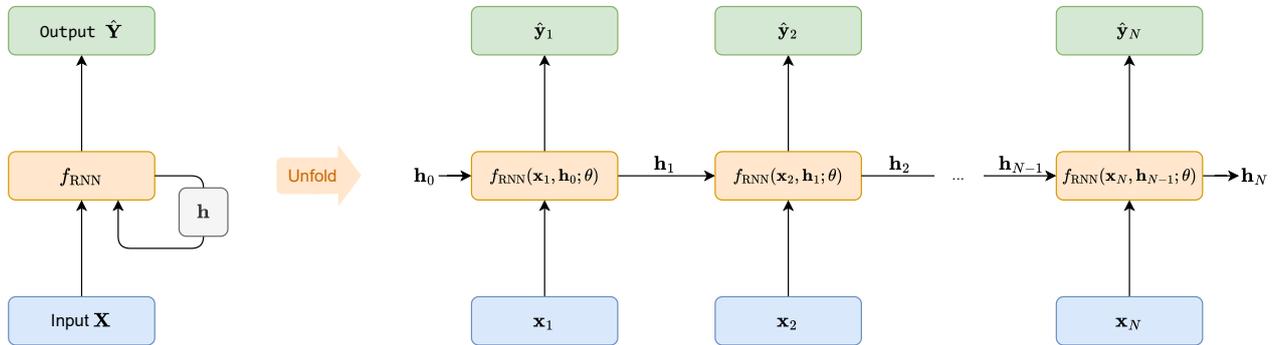


FIGURE 2.5: An illustration of a recurrent neural network architecture with the unfolded view to show the processing of each time step (token in the sequence). Each RNN cell,  $f_{\text{RNN}}(\mathbf{x}_t, \mathbf{h}_{t-1}; \theta)$ , represents the operation specified in Equations 2.44 and 2.45.

### The Vanishing Gradient Problem

RNNs are trained using backpropagation through time (BPTT), where gradients are propagated not only across layers but also over multiple time steps. This leads to the well-known challenge of the *vanishing (or exploding) gradient problem* caused primarily by the multiplication of Jacobians through time steps:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{h}_t} = \prod_{k=t}^T \frac{\partial \mathbf{h}_k}{\partial \mathbf{h}_{k-1}} \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{h}_N}, \quad (2.47)$$

where  $\mathcal{L}$  denotes the loss, and  $\mathbf{h}_t$  the hidden representation at time step  $t$  in a total number of  $T$  steps.

Similar to the case of FNNs, the gradient vanishes if the eigenvalues of the Jacobian are less than one, or explodes otherwise. These difficulties severely limit the ability of simple RNNs to learn long-term dependencies.

## Encoder-decoder Architecture of RNN

A simple RNN is designed either to predict the next token in a sequence or to map an input sequence to a fixed output for classification, but it cannot directly generate an output sequence of variable length conditioned on an input sequence. The *encoder-decoder architecture* was introduced by (Cho et al., 2014) and (Sutskever et al., 2014) to overcome the limitations of standard RNN language models in handling sequence-to-sequence tasks such as machine translation and dialogue response generation. In this framework, an **encoder** RNN reads an input sequence  $(x_1, \dots, x_N)$  and maps it to a sequence of hidden states  $(h_1, \dots, h_N)$ . A fixed-length representation  $h_N$  is then passed to a **decoder** RNN, which generates an output sequence  $(\hat{y}_1, \dots, \hat{y}_M)$  autoregressively. The beginning and ending of the generated sequence are marked as special tokens, [BOS] and [EOS]. To address issues of vanishing gradients and long-range dependency learning in standard RNNs, variants such as the Gated Recurrent Unit (GRU, Chung et al. 2014) and Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber 1997) have been developed. These gated architectures allow the network to retain information over longer sequences, improving performance on tasks like machine translation and dialogue response generation. Figure 2.6 depicts the architecture of a typical encoder-decoder RNN.

Formally, the encoder computes:

$$h_t = f_{\text{enc}}(h_{t-1}, x_t; \theta_{\text{enc}}), \quad (2.48)$$

while the decoder computes:

$$\hat{y}_t, s_t = g_{\text{dec}}(s_{t-1}, \hat{y}_{t-1}; \theta_{\text{dec}}), \quad (2.49)$$

where  $f_{\text{enc}}$  and  $g_{\text{dec}}$  are recurrent functions (e.g., RNN, GRU, or LSTM),  $\{x_t, \hat{y}_{t-1}\}$  are embeddings of source and generated tokens,  $s_t$  is the hidden state at decoding time  $t$ .

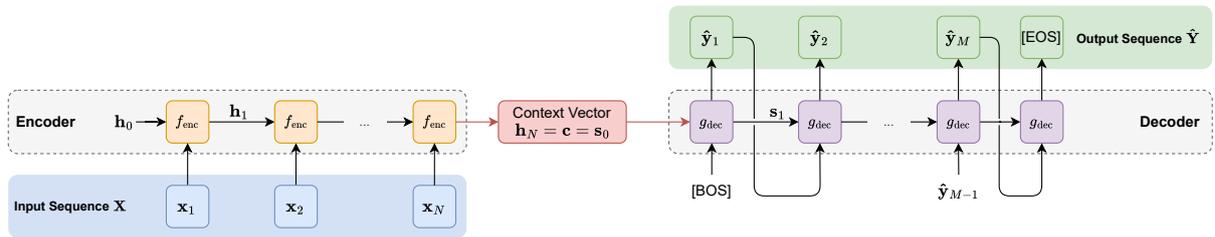


FIGURE 2.6: An illustration of an encoder-decoder recurrent neural network architecture. The hidden state of the last input token  $h_N$  is used as the context vector, which is used as the first hidden state  $s_0$  in the decoder to initialise the generation.

## Attention Mechanisms in RNNs

While RNNs and their variants (e.g., GRU, LSTM) can model sequential data, they often struggle with very long sequences due to vanishing gradients and limited memory capacity. In the vanilla encoder-decoder architecture as illustrated in Figure 2.6, the context is limited to the last encoder state, limiting the performance on long sequences. To overcome this, *attention mechanisms* were introduced to augment RNNs by allowing the model to dynamically focus on relevant parts of the input sequence when producing each output (Bahdanau, 2014).

Formally, given a decoder hidden state  $s_m$  at decoder time step  $m$  and encoder hidden states  $(h_1, h_2, \dots, h_N)$ , the attention score  $e_{n,m}$  for encoder position  $n$  is computed with a scoring function parameterised with  $\theta_{\text{score}}$ :

$$e_{n,m} = f_{\text{score}}(h_n, s_m; \theta_{\text{score}}). \quad (2.50)$$

The attention weights are normalised with a softmax function:

$$\alpha_{n,i} = \frac{\exp(e_{n,m})}{\sum_{j=1}^N \exp(e_{j,m})}, \quad (2.51)$$

and the context vector is obtained as a weighted sum of encoder states:

$$\mathbf{c}_m = \sum_{n=1}^N \alpha_{n,m} \mathbf{h}_n. \quad (2.52)$$

The context  $\mathbf{c}_m$  is then combined with the decoder hidden state to produce the output:

$$\hat{\mathbf{y}}_{m+1}, \mathbf{s}_{m+1} = g_{\text{dec}}(\mathbf{s}_m, \mathbf{c}_m; \boldsymbol{\theta}_{\text{dec}}), \quad (2.53)$$

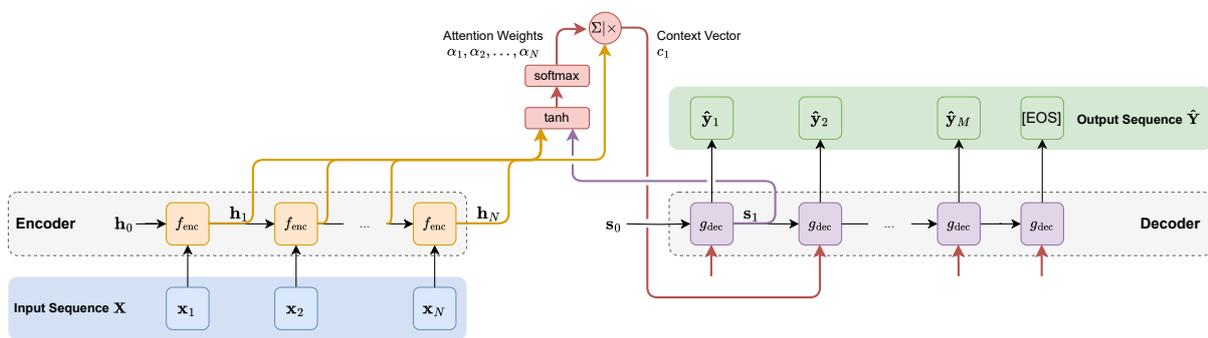


FIGURE 2.7: An illustration of an implementation of the attention mechanism in an encoder-decoder recurrent neural network architecture. The attention scoring function in Equation 2.50 is implemented as  $f_{\text{score}}(\mathbf{h}_n, \mathbf{s}_m; \boldsymbol{\theta}_{\text{score}}) = \tanh(\mathbf{W}_x \mathbf{h}_n + \mathbf{W}_y \mathbf{s}_m + \mathbf{b}_{\text{score}})$  where trainable parameters  $\boldsymbol{\theta}_{\text{score}}$  are weight matrices  $\mathbf{W}_x, \mathbf{W}_y$  and bias vector  $\mathbf{b}_{\text{score}}$ . Attention scores of all encoder hidden states are normalised with the softmax function, following a weighted sum operation to obtain the context vector  $\mathbf{c}_m$  (Equation 2.52). In the illustration,  $m = 1$ .

This mechanism allows the decoder to directly access information from any part of the source sequence, improving the modelling of long-range dependencies and enabling explicit alignment between input and output tokens. Attention mechanisms also inspire the development of Transformer architectures, which we later discuss in Section 2.5.4.

## 2.5.4 Transformers

The Transformer architecture (Vaswani et al., 2017) represents a paradigm shift in sequence modelling. Unlike recurrent or convolutional architectures, Transformers dispense with recurrence entirely and rely solely on *attention* mechanisms to capture dependencies between tokens. This design enables parallelisation during training and effectively models both short- and long-range dependencies.

### Attention Mechanism in Transformers

The fundamental building block of Transformers is *scaled dot-product attention*. It takes three input matrices: queries, keys, and values. Later in Section 2.5.4, we examine two typical input set-ups of the attention mechanism illustrated in Figure 2.8 for building Transformer-based language models.

The query, key, and value matrices are first linearly projected with weight matrices  $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ . Then, it computes attention weights as a similarity between queries and keys, scaled by the factor

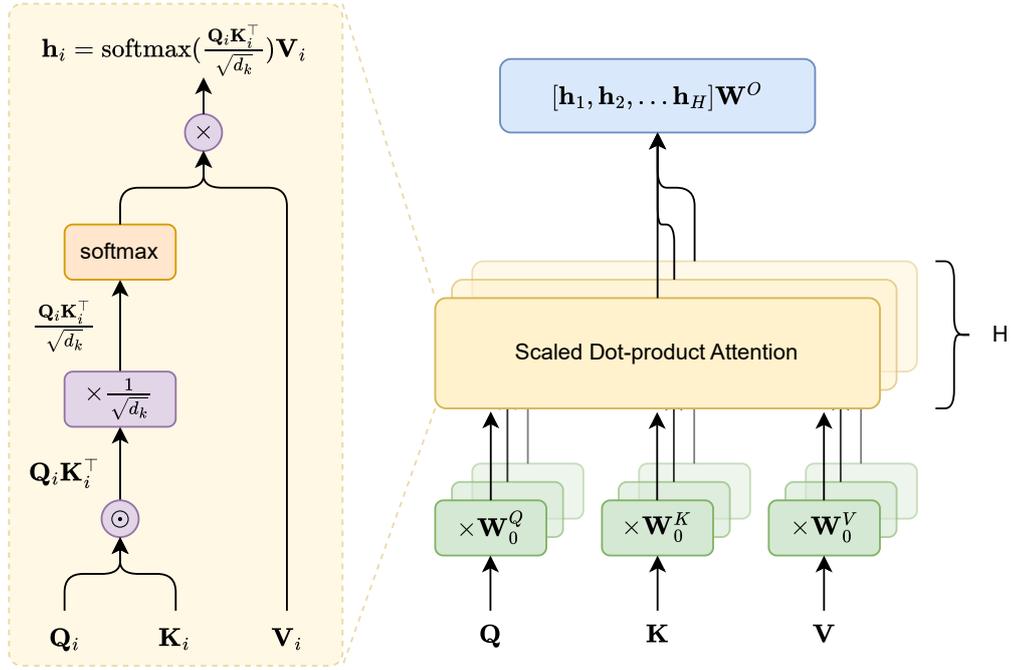


FIGURE 2.8: An illustration of the scaled dot-product attention in a multi-head attention set-up.

$\sqrt{d_k}$  to control variance:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2.54)$$

with  $d_k$  as the dimensionality of the queries and keys.

The scaling factor  $1/\sqrt{d_k}$  is crucial for numerical stability during training. The variance of the dot products  $QK^T$  increases proportionally to  $d_k$ . Large logits drive the softmax function into regions with very small gradients, leading to vanishing gradients and poor learning dynamics. By dividing by  $\sqrt{d_k}$ , the variance of the dot products is normalised, keeping the softmax in a regime where it is more sensitive to differences between queries and keys.

### Multi-Head Attention (MHA)

In practice, rather than computing attention once, the Transformer employs multiple attention “heads”, each with different learned projections, allowing the model to jointly capture diverse relationships. Specifically,

$$h_i = \text{Attention}(Q_i, K_i, V_i), \quad i = 1, \dots, H, \quad (2.55)$$

$$f_{\text{MHA}}(X) = [h_1; h_2; \dots; h_H]W^O, \quad (2.56)$$

where  $H$  is the number of heads,  $[\cdot]$  denotes concatenation, and  $W^O$  is a learnable projection matrix. This design allows the model to attend simultaneously to information at multiple positions and representation subspaces.

### Transformer Block

A Transformer is built from many blocks, which contain more than merely Multi-Head Attention (MHA) modules. The block consists of the following core components: a multi-head attention (MHA) module (as discussed in this section) and a position-wise FNN (discussed in Section 2.5.1), each wrapped with a residual connection (Section 2.5.2) and layer normalisation (Section 2.4.3). As

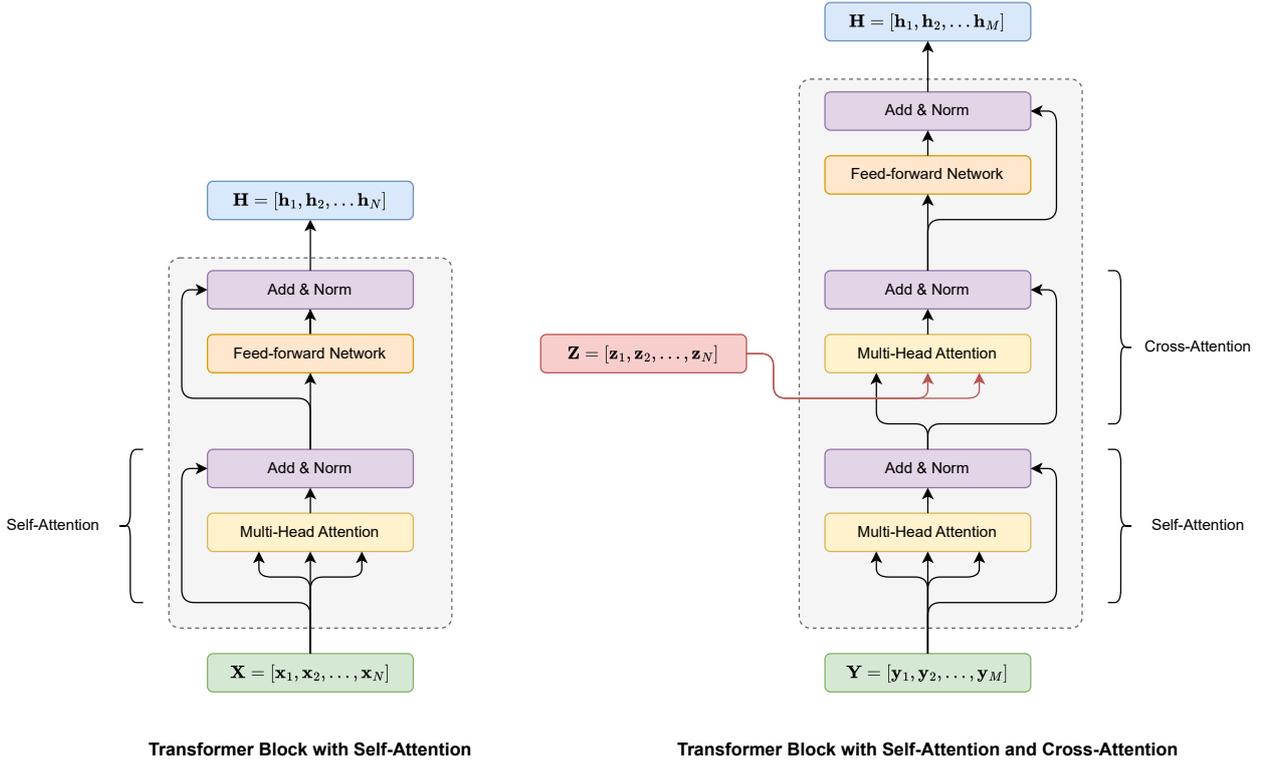


FIGURE 2.9: Left: Transformer block with self-attention takes a sequence of vectors as input stacked in a matrix  $\mathbf{X}$  and outputs a sequence of hidden vectors stacked in a matrix  $\mathbf{H}$ . The Transformer block starts with a multi-head attention (MHA), which takes  $\mathbf{X}$  as queries, keys and values, followed by a residual connection (Add operation) and layer normalisation (Norm). The output of MHA is fed into a FNN with a non-linear activation function. Right: the Transformer block has an additional cross-attention mechanism, which takes an additional sequence of vectors stacked in a matrix  $\mathbf{Z}$  as keys and values while taking the queries from the output of the previous self-attention module.

illustrated in Figure 2.9, the Transformer block typically has two setups: with self-attention only and with cross-attention in addition.

**Transformer Block with Self-Attention** takes as input a matrix  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$ , where  $N$  is the sequence length and  $d$  the hidden dimension, and produces an output matrix  $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_N] \in \mathbb{R}^{N \times d}$ .

In the self-attention module, queries, keys, and values are all derived from the input:  $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{X}$ . Therefore, the mechanism here is referred to as *self-attention*. After applying MHA, a residual connection and layer normalisation are used:

$$\mathbf{H}' = \text{Norm}(\mathbf{X} + \text{MHA}(\mathbf{X}, \mathbf{X}, \mathbf{X})) \in \mathbb{R}^{N \times d}. \quad (2.57)$$

Each row vector of  $\mathbf{H}'$  is then passed through a position-wise FNN, which consists of two linear transformations separated by a non-linearity:

$$\text{FNN}(\mathbf{h}) = \mathbf{W}_2 \phi_{\text{non-linear}}(\mathbf{W}_1 \mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2, \quad (2.58)$$

where  $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d}$ ,  $\mathbf{W}_2 \in \mathbb{R}^{d \times d_1}$ ,  $\mathbf{b}_1 \in \mathbb{R}^{d_1}$ ,  $\mathbf{b}_2 \in \mathbb{R}^d$ . The hidden dimension  $d_1$  is typically set larger than  $d$  (e.g.  $d_1 = 4d$ ) to increase representational capacity. A second residual connection and layer

normalisation are then applied:

$$\mathbf{H} = \text{Norm}(\mathbf{H}' + \text{FNN}(\mathbf{H}')) \in \mathbb{R}^{N \times d}. \quad (2.59)$$

**Transformer Block with Self-Attention and Cross-Attention** takes as input matrices  $\mathbf{Y} = [y_1, \dots, y_M] \in \mathbb{R}^{M \times d}$  and  $\mathbf{Z} = [z_1, \dots, z_N] \in \mathbb{R}^{N \times d}$ , where  $M$  and  $N$  are sequence lengths and  $d$  the hidden dimension, and produces an output matrix  $\mathbf{H} = [h_1, \dots, h_N] \in \mathbb{R}^{M \times d}$ .

The matrix  $\mathbf{Y}$  is first fed into the multi-head attention (MHA) module in the self-attention fashion,  $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{Y}$ , that we have just discussed, followed by a residual connection and layer normalisation operation:

$$\tilde{\mathbf{H}} = \text{Norm}(\mathbf{Y} + \text{MHA}(\mathbf{Y}, \mathbf{Y}, \mathbf{Y})) \in \mathbb{R}^{M \times d}. \quad (2.60)$$

A cross-attention MHA module follows and takes  $\mathbf{Q} = \tilde{\mathbf{H}}$  and  $\mathbf{K} = \mathbf{V} = \mathbf{Z}$ .

$$\mathbf{H}' = \text{Norm}(\tilde{\mathbf{H}} + \text{MHA}(\tilde{\mathbf{H}}, \mathbf{Z}, \mathbf{Z})) \in \mathbb{R}^{M \times d}. \quad (2.61)$$

Like the Transformer block with only self-attention,  $\mathbf{H}'$  is fed into a FNN with a residual connection, followed by layer normalisation.

$$\mathbf{H} = \text{Norm}(\mathbf{H}' + \text{FNN}(\mathbf{H}')) \in \mathbb{R}^{M \times d}. \quad (2.62)$$

In summary, a Transformer block processes a sequence of vectors in parallel through the self-attention mechanism via efficient matrix-matrix multiplication. Compared to recurrent neural networks, self-attention enables more effective modelling of long-range dependencies between input elements without sequential computation. The block further incorporates a FNN to introduce non-linear transformations and employs residual connections and layer normalisation to stabilise and accelerate training.

Transformer blocks can be stacked, analogous to the layers of a deep FNN. Each block maintains its own set of parameters for the multi-head attention module, the FNN, and normalisation layers. The output of one block serves as the input to the next. Importantly, the input and output dimensions of each block are aligned by design, a property that ensures the residual connections are well-defined.

## Positional Encoding

Since the attention mechanism in Transformers is permutation-invariant, additional mechanisms are required to encode order, which is essential for text processing. This is achieved through sinusoidal *positional encodings* (PEs), which are added to the input embeddings to inject sequence position information (Vaswani et al., 2017). For a position  $\text{pos}$  and dimension  $i$ , these are defined as:

$$\text{PE}_{(\text{pos}, 2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d}}\right), \quad \text{PE}_{(\text{pos}, 2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/d}}\right). \quad (2.63)$$

The use of sine and cosine at varying frequencies ensures that positions have unique encodings and that the model can generalise to sequences longer than those seen during training. This is illustrated in Figure 2.10. In later variants, learned positional embeddings have also been adopted (Devlin et al., 2019; Radford et al., 2018).

## The Encoder-decoder Architecture of Transformers

The Transformer model was first proposed for sequence-to-sequence modelling in translation. Analogue to recurrent neural networks, the Transformer also employs an encoder-decoder architecture.

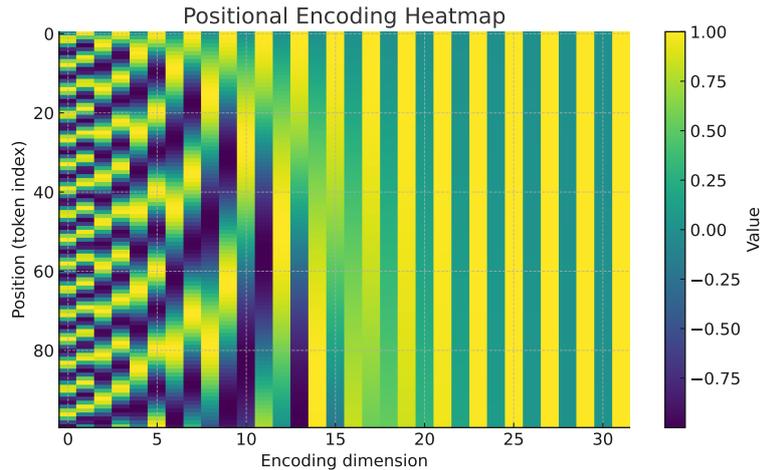


FIGURE 2.10: A visualisation of positional encoding values across different positions and dimensions.

**The Transformer Encoder** is composed of  $L_{\text{enc}}$  Transformer blocks with self-attention. The input vector sequence  $\mathbf{X} = [x_1, \dots, x_N] \in \mathbb{R}^{N \times d}$  is first added with positional encodings as explained in Section 2.5.4. Then, the encoder produces a sequence of contextualised representations  $\mathbf{Z} = [z_1, \dots, z_N] \in \mathbb{R}^{N \times d}$ .

**The Transformer Decoder** aims to take the encoded features  $\mathbf{Z}$  to produce the output sequence  $[\hat{y}_1, \dots, \hat{y}_M] \in \mathbb{R}^{M \times d}$ . It is composed of  $L_{\text{dec}}$  Transformer blocks with self-attention and cross-attention. The decoding procedure is auto-regressive: at each time  $j = 1, \dots, M$ , the input to the Transformer decoder is a matrix  $\hat{\mathbf{Y}} = [\hat{y}_0, \dots, \hat{y}_{j-1}] \in \mathbb{R}^{j \times d}$ . This matrix contains the feature vectors of the previously generated predictions  $\hat{y}_0, \dots, \hat{y}_{j-1}$ . The initial vector  $\hat{y}_0$  is typically chosen as either a zero vector or a special learned embedding that signals the start of the decoding process (e.g., a [BOS] token). Same as the encoder, position encodings are added to the generated sequence before decoding for the new token. This sequence, together with the encoder output matrix  $\mathbf{Z}$ , is passed to a stack of Transformer blocks with both self-attention and cross-attention. Note that the self-attention in the decoder is masked so that the contextualised representation of an element in the sequence only attends to earlier ones. The result is a hidden representation  $\hat{\mathbf{H}} = [\hat{h}_0, \dots, \hat{h}_{j-1}]$ . Finally, the most recent hidden state  $\hat{h}_{j-1}$  is projected through a linear transformation followed by a softmax activation to obtain a probability distribution over the vocabulary:

$$\hat{y}_j^{\text{out}} = \phi_{\text{softmax}}(\mathbf{W}\hat{h}_{j-1} + \mathbf{b}), \quad (2.64)$$

where  $\mathbf{W} \in \mathbb{R}^{|V| \times d}$  and  $\mathbf{b} \in \mathbb{R}^{|V|}$  are trainable parameters, and  $|V|$  is the vocabulary size. An output token  $\hat{y}_j$  is then sampled (or selected via greedy decoding or beam search), and the process iterates until an end-of-sequence token is generated or the maximum length is reached.

## 2.6 Supervised Learning

Supervised learning constitutes one of the most fundamental paradigms in machine learning. It is characterised by the availability of a dataset consisting of input-output pairs, from which a model learns a mapping that generalises to unseen data. This paradigm has been central to advances in NLP, enabling the development of models for tasks such as classification, regression, and sequence-to-sequence learning.

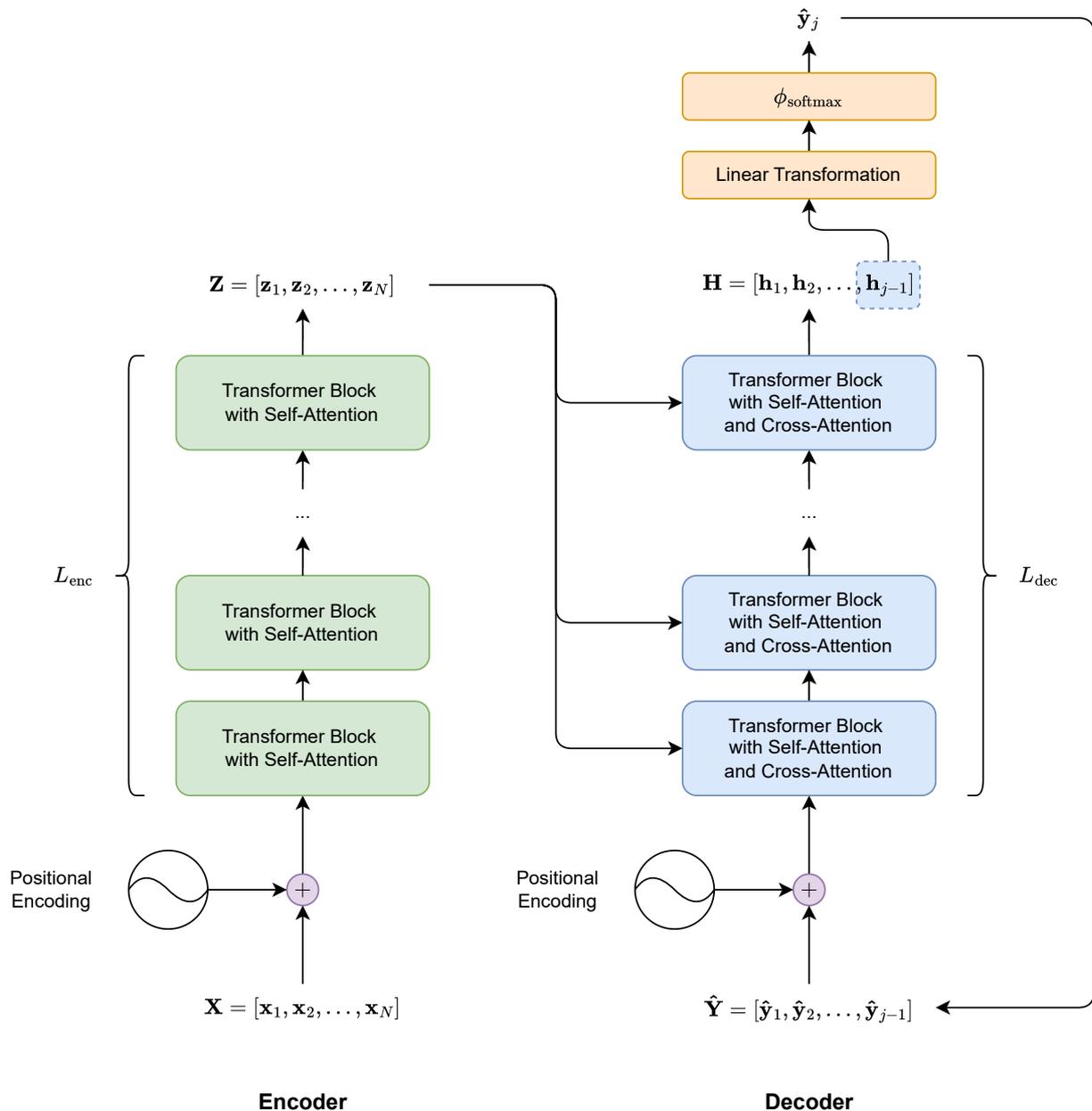


FIGURE 2.11: An illustration of the Transformer-encoder-decoder architecture. Left: the Transformer encoder is composed of a stack of  $L_{\text{enc}}$  Transformer blocks with self-attention. Right: the Transformer decoder is composed of a stack of  $L_{\text{dec}}$  Transformer blocks with self-attention and cross-attention, followed by a linear transformation and a softmax activation function, and auto-regressively produces the next element. The positional encodings are added to the input for both the encoder and the decoder.

### 2.6.1 Formal Definition

Let  $D = \{(x_i, y_i)\}_{i=1}^N$  be a training dataset of  $N$  examples, where each  $x_i \in \mathcal{X}$  denotes an input (e.g., a text sequence or feature vector) and  $y_i \in \mathcal{Y}$  is the corresponding target label. The goal of supervised learning is to learn a function

$$f_{\theta} : \mathcal{X} \rightarrow \mathcal{Y}, \quad (2.65)$$

parameterised by  $\theta$ , that minimises the expected risk

$$R(\theta) = \mathbb{E}_{(x,y) \sim p_{\text{data}}} [\ell(f_{\theta}(x), y)], \quad (2.66)$$

where  $\ell(\cdot, \cdot)$  is a task-specific loss function and  $p_{\text{data}}$  is the (unknown) data distribution. Since  $p_{\text{data}}$  is inaccessible, optimisation proceeds via empirical risk minimisation (ERM):

$$\hat{R}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i). \quad (2.67)$$

The minimisation of  $\hat{R}(\boldsymbol{\theta})$  is typically performed using gradient-based optimisation discussed in Section 2.4.2.

### 2.6.2 Classification

In classification task, the label space is discrete,  $\mathcal{Y} = \{1, \dots, C\}$ , where  $C$  is the number of categories. The model produces a probability distribution over classes via the softmax function:

$$p_{\boldsymbol{\theta}}(y = c \mid \mathbf{x}) = \frac{\exp(\mathbf{w}_c^\top \mathbf{h}_x + \mathbf{b}_c)}{\sum_{c'=1}^C \exp(\mathbf{w}_{c'}^\top \mathbf{h}_x + \mathbf{b}_{c'})}, \quad (2.68)$$

where  $\mathbf{h}_x$  denotes the hidden representation of input  $\mathbf{x}$ , and  $\mathbf{w}_c, \mathbf{b}_c$  are class-specific parameters.

The standard objective is the cross-entropy loss as we have discussed in Section 2.4.1. Classification underpins numerous NLP applications, including emotion recognition and intent categorisation.

### 2.6.3 Regression

In regression tasks, the label space is continuous,  $\mathcal{Y} \subseteq \mathbb{R}$ . The model predicts a real-valued output

$$\hat{y}_i = f_{\boldsymbol{\theta}}(\mathbf{x}_i), \quad (2.69)$$

and performance is often measured using the mean squared error (MSE) as we have discussed in Section 2.4.1. Regression is used in NLP for tasks such as readability estimation, semantic similarity scoring, and predicting continuous affective dimensions.

### 2.6.4 Sequence-to-Sequence Learning

A more complex supervised paradigm involves sequence-to-sequence (seq2seq) learning, where both the input and output are variable-length sequences. This framework generalises standard classification: at each step, the model predicts the next token from a finite vocabulary, effectively performing a classification over the vocabulary space conditioned on the source sequence and previous target tokens.

Formally, given an input sequence of length  $T$

$$\mathbf{x} = (w_1, \dots, w_T) \quad (2.70)$$

and a target sequence of length  $T'$

$$\mathbf{y} = (y_1, \dots, y_{T'}). \quad (2.71)$$

The objective is to model the conditional distribution

$$p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T'} p_{\boldsymbol{\theta}}(y_t \mid \mathbf{y}_{<t}, \mathbf{x}), \quad (2.72)$$

where  $\mathbf{y}_{<t} = (y_1, \dots, y_{t-1})$  denotes the prefix of the target sequence.

The training objective minimises the negative log-likelihood (NLL):

$$L(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T_i} \log p_{\boldsymbol{\theta}}(y_{i,t} \mid \mathbf{y}_{i,<t}, \mathbf{x}_i). \quad (2.73)$$

Seq2seq models thus extend the classification paradigm to autoregressive prediction, where a sequence of classification decisions generates a structured output. They became central to NLP tasks such as machine translation (Bahdanau, 2014; Vaswani et al., 2017), summarisation, and dialogue response generation.

### 2.6.5 Advantages and Limitations

Supervised learning offers several advantages. First, it provides conceptual clarity: the problem formulation is mathematically well-defined and empirically effective. Second, it is broadly applicable: the paradigm supports diverse tasks, including classification, regression, and sequence-to-sequence modelling. Finally, it offers strong empirical performance: with sufficiently large and representative training data, supervised models often achieve state-of-the-art results.

However, the approach also faces limitations. Supervised learning is heavily data-dependent, and constructing high-quality labelled datasets is expensive and time-intensive. Models may also suffer from generalisation gaps, performing poorly under domain shift. Furthermore, label noise introduced during human annotation can degrade performance, particularly in subjective tasks such as emotion recognition.

Despite these challenges, supervised learning remains the foundation of many machine learning applications and provides a critical basis for more advanced paradigms.

## 2.7 Transfer Learning

Supervised learning, despite its effectiveness, is limited by the availability of large annotated datasets. In many practical scenarios, especially in specialised domains, only a small quantity of labelled data is accessible. Transfer learning addresses this challenge by leveraging knowledge acquired in one setting (the source domain or task) to improve learning in another (the target domain or task). This paradigm has become central to natural language processing, particularly following the success of large-scale pre-trained language models.

Formally, let a domain be defined as

$$D = (\mathcal{X}, P(\mathcal{X})), \quad (2.74)$$

where  $\mathcal{X}$  is the feature space and  $P(\mathcal{X})$  is the marginal distribution over inputs. A task is defined as

$$T = (\mathcal{Y}, f(\cdot)), \quad (2.75)$$

where  $\mathcal{Y}$  is the label space and  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is the predictive function. Given a source domain  $D_S$  with task  $T_S$ , and a target domain  $D_T$  with task  $T_T$ , transfer learning seeks to improve the learning of  $f_T$  in  $(D_T, T_T)$  using information from  $(D_S, T_S)$ , where typically  $D_S \neq D_T$  and/or  $T_S \neq T_T$  (Pan and Yang, 2009).

Several complementary strategies for transfer learning have emerged in NLP, which we discuss in the following subsections.

### 2.7.1 Domain Adaptation

Domain adaptation addresses the case where the source and target tasks are identical ( $T_S = T_T$ ) but the input distributions differ ( $P_S(\mathcal{X}) \neq P_T(\mathcal{X})$ ). A sentiment classifier trained on product reviews, for example, may not generalise to movie reviews due to distributional differences.

One common approach is to *fine-tune pre-trained models*. Large language models pretrained on corpora such as Wikipedia or Common Crawl capture broad linguistic regularities, and can be adapted to target domains with small amounts of labelled data (Devlin et al., 2019; Raffel et al., 2020). Another strategy is *instance reweighting*, where source-domain examples are weighted to better reflect the target distribution (Jiang and Zhai, 2007). *Feature alignment* methods instead encourage the model to learn domain-invariant representations, often using adversarial training (Ganin et al., 2016), where a discriminator distinguishes domains while the encoder seeks to confuse it. Finally, *unsupervised domain adaptation* leverages unlabelled target-domain data, for example, through self-training with pseudo-labels or consistency regularisation (French et al., 2018). Domain adaptation has proven especially impactful in NLP due to the abundance of unlabelled text in specialised domains.

### 2.7.2 Data Augmentation

Data augmentation seeks to artificially expand the training dataset to improve generalisation.

Formally, given a dataset  $D = \{(x_i, y_i)\}_{i=1}^N$ , augmentation applies a stochastic transformation  $\tau : \mathcal{X} \rightarrow \mathcal{X}$  to inputs, yielding

$$D' = \{(\tau(x_i), y_i)\}_{i=1}^N. \quad (2.76)$$

Training on  $D \cup D'$  reduces overfitting and improves robustness to linguistic variability.

In NLP, augmentation methods mitigate the scarcity of labelled data by introducing lexical and structural variability. *Lexical substitution* replaces words with synonyms or semantically similar alternatives (Zhang et al., 2015). *Back-translation* generates paraphrases by translating text into another language and back again (Sennrich et al., 2016). Another approach is *noise injection*, where tokens are randomly deleted, shuffled, or masked to encourage robustness (Xie et al., 2020a). *Template-based generation* creates synthetic examples with rules or neural templates (Fadaee et al., 2017), while large language models have recently been used for augmentation by generating paraphrases, counterfactuals, or task-specific examples.

### 2.7.3 Multitask-Learning

Another form of transfer arises in *multi-task learning* (MTL), where a model is trained jointly on multiple related tasks to improve generalisation through shared representations (Caruana, 1997). For example, part-of-speech tagging, syntactic parsing, and semantic role labelling can be trained together using a shared encoder, encouraging inductive transfer between tasks.

Formally, given tasks  $\{T_1, \dots, T_K\}$ , the joint training objective is

$$\mathcal{L}(\theta) = \sum_{k=1}^K \lambda_k \mathcal{L}_k(\theta), \quad (2.77)$$

where  $\mathcal{L}_k(\theta)$  is the loss for task  $T_k$  and  $\lambda_k$  is a weighting factor. MTL has shown consistent benefits in NLP by enabling models to learn linguistically meaningful shared structures (Williams et al., 2018).

### 2.7.4 Parameter Efficient Fine-tuning

Recent advances in transfer learning have emphasised *parameter-efficient fine-tuning*, which adapts large pre-trained models without updating all parameters. Instead of modifying the full parameter set  $\theta$  of a pre-trained model, a small set of additional parameters  $\Delta\theta$  is introduced and optimised,

while the original weights remain frozen. Formally, given a pre-trained model  $f_{\theta}$ , parameter-efficient methods learn

$$f_{\theta, \Delta\theta}(x) \quad \text{with} \quad \theta \text{ frozen, } \Delta\theta \text{ trainable.} \quad (2.78)$$

**Adapters** One approach is the use of *adapters* (Houlsby et al., 2019), small trainable modules inserted between frozen pre-trained layers. If  $\mathbf{h}$  denotes the hidden representation of a layer, an adapter typically applies a bottleneck transformation:

$$\mathbf{h}' = \mathbf{h} + \mathbf{W}_{\text{up}} \sigma(\mathbf{W}_{\text{down}} \mathbf{h}), \quad (2.79)$$

where  $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r \times d}$  projects the  $d$ -dimensional hidden state into a lower-dimensional space  $r \ll d$ ,  $\sigma$  is a non-linear activation function, and  $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d \times r}$  projects it back. Only  $\mathbf{W}_{\text{up}}$  and  $\mathbf{W}_{\text{down}}$  are updated during fine-tuning, leading to a significant reduction in trainable parameters when adapting to a new task.

**Low-Rank Adaptation (LoRA)** Another widely adopted method is *low-rank adaptation (LoRA)* (Hu et al., 2022). Instead of updating a full weight matrix  $\mathbf{W} \in \mathbb{R}^{d \times k}$ , LoRA constrains the update to a low-rank decomposition:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}, \quad \text{with} \quad \Delta\mathbf{W} = \mathbf{A}\mathbf{B}, \quad (2.80)$$

where  $\mathbf{A} \in \mathbb{R}^{d \times r}$  and  $\mathbf{B} \in \mathbb{R}^{r \times k}$  with  $\text{rank } r \ll \min(d, k)$ . This reduces the number of trainable parameters from  $\mathcal{O}(dk)$  to  $\mathcal{O}(r(d+k))$ , while retaining much of the expressive power of full-rank updates.

**Prefix Tuning** A third strategy is *prefix tuning* (Li and Liang, 2021), which modifies the input to each Transformer layer rather than its weights. For a self-attention mechanism with key and value matrices  $(\mathbf{K}, \mathbf{V})$ , prefix tuning introduces trainable prefix vectors  $(\mathbf{K}_p, \mathbf{V}_p)$  of length  $m$ , which are concatenated with the original key and value sequences:

$$\mathbf{K}' = [\mathbf{K}_p; \mathbf{K}], \quad \mathbf{V}' = [\mathbf{V}_p; \mathbf{V}]. \quad (2.81)$$

During training, only  $(\mathbf{K}_p, \mathbf{V}_p)$  are updated, while the pre-trained model parameters remain frozen. This allows the model's behaviour to be steered towards the target task with minimal parameter updates.

### 2.7.5 Advantages and Limitations

Transfer learning has become a cornerstone of modern NLP due to several key advantages. First, it enables effective learning in settings where annotated data is scarce. By leveraging knowledge encoded in pre-trained models or related tasks, transfer learning reduces the dependence on costly human-labelled corpora. Second, it provides strong empirical performance across a wide range of downstream tasks, from classification and sequence labelling to text generation. Large-scale pre-training captures broad linguistic regularities, enabling models to generalise well even in specialised domains. Third, transfer learning often accelerates convergence during fine-tuning: rather than optimising models from scratch, training can begin from a strong initialisation, which improves stability and reduces computational cost. Finally, parameter-efficient fine-tuning methods extend these benefits by making adaptation feasible for very large models, lowering memory and training overhead while still retaining task-specific flexibility.

Despite these strengths, transfer learning also faces notable limitations. The effectiveness of knowledge transfer can depend heavily on the similarity between source and target domains; when the domains diverge significantly, negative transfer may occur, leading to degraded performance. Moreover, while large-scale pre-trained models enable strong results, they require vast computational

resources for pretraining, raising concerns about accessibility and environmental sustainability. Another limitation lies in adaptability: pre-trained models may encode unwanted biases present in their training data, and naive transfer can propagate or even amplify these biases in downstream tasks. Finally, although parameter-efficient methods reduce the cost of fine-tuning, they still rely on storing multiple task-specific parameter sets, which can be burdensome in large multi-task scenarios.

In summary, transfer learning provides a powerful paradigm that mitigates data scarcity and enhances performance across tasks, but challenges remain in ensuring robustness, fairness, and efficiency when adapting to new domains and applications.

## 2.8 Unsupervised Learning

Unsupervised learning (also termed as self-supervised learning) refers to the paradigm in which models learn structure from unlabelled data without explicit supervision. In natural language processing, this approach is particularly important given the abundance of raw text and the relative scarcity of annotated corpora. By discovering patterns, latent structures, and predictive regularities in unlabelled text, unsupervised learning forms the foundation of modern representation learning and large-scale pretraining.

### 2.8.1 Language Modelling

One of the most influential unsupervised learning tasks in natural language processing is language modelling, where the task is to model the probability distribution over sequences of tokens.

#### Autoregressive Language Models

Given a sequence of  $T$  tokens  $x = (w_1, w_2, \dots, w_T)$ , the joint probability is factorised autoregressively as:

$$p(w_1, w_2, \dots, w_T) = \prod_{t=1}^T p(w_t | \mathbf{w}_{<t}), \quad (2.82)$$

where  $\mathbf{w}_{<t} = (w_1, \dots, w_{t-1})$

Training proceeds by maximising the log-likelihood over the dataset  $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$ :

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^{T^{(i)}} \log p_{\theta}(w_t^{(i)} | \mathbf{w}_{<t}^{(i)}). \quad (2.83)$$

This objective encourages the model to learn contextual dependencies in language, forming the basis for generation tasks.

#### Masked Language Models

An alternative is masked language modelling (MLM), introduced by BERT (Devlin et al., 2019). Here, tokens are randomly masked, and the model is trained to predict them given their context:

$$\mathcal{L}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \sum_{t \in M(x)} \log p_{\theta}(w_t | \mathbf{x}_{\setminus t}), \quad (2.84)$$

where  $M(x)$  is the set of masked positions, and  $\mathbf{x}_{\setminus t}$  is the sequence with  $w_t$  token masked out.

MLM encourages bidirectional contextual representations, which are useful for representation learning.

## 2.8.2 Advantages and Limitations

Unsupervised learning offers several advantages for NLP. The most significant is scalability: vast quantities of raw text can be exploited without the need for costly human annotation. This enables models to capture broad linguistic patterns, world knowledge, and domain-specific structures that are difficult to obtain from limited labelled corpora. Furthermore, unsupervised pretraining often serves as an effective initialisation, substantially improving performance on downstream supervised tasks through fine-tuning or transfer learning.

However, unsupervised learning also has limitations. Objectives such as language modelling may encourage the memorisation of surface regularities rather than the acquisition of deeper semantics. Moreover, the absence of explicit supervision makes evaluation more challenging: it is often unclear whether the representations learned align with the target task requirements. Unsupervised training also demands significant computational resources, particularly for LLMs, raising concerns about efficiency and accessibility. Finally, since unsupervised methods rely on naturally occurring text, they may inherit biases and artefacts from the data, which can propagate to downstream applications.

In summary, while unsupervised learning provides the foundation for modern NLP, its success depends on careful objective design, large-scale computation, and subsequent fine-tuning to align representations with task-specific needs.

## 2.9 Semi-Supervised Learning

Semi-supervised learning is a paradigm in which models are trained on a combination of a small labelled dataset, as in supervised learning, and a larger unlabelled dataset, as in unsupervised learning. This approach leverages the abundant unlabelled text to improve model performance, while still guiding learning with task-specific supervision. Semi-supervised learning is particularly valuable in NLP (Dai and Le, 2015), where obtaining high-quality annotations can be costly and time-consuming, but large volumes of raw text are readily available.

### 2.9.1 Semi-Supervised Learning Techniques

**Self-training** is one of the most fundamental and well-established approaches in semi-supervised NLP. In self-training, a model is first trained on the labelled data, then used to generate pseudo-labels for the unlabelled data (Yarowsky, 1995). These pseudo-labelled examples are then incorporated into training, typically with confidence-based filtering:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{labelled}}(\theta) + \lambda \mathcal{L}_{\text{pseudo}}(\theta), \quad (2.85)$$

where  $\lambda$  controls the contribution of pseudo-labelled data. Self-training iteratively improves model predictions on unlabelled data and can significantly expand the effective training set.

Beyond self-training, several other semi-supervised learning techniques have been proposed. **Consistency regularisation** (Xie et al., 2020b) encourages models to produce stable predictions under input perturbations, dropout noise, or data augmentation, thereby promoting smoother decision boundaries. **Co-training** trains multiple models or views of the data and allows them to teach each other by exchanging high-confidence predictions (Blum and Mitchell, 1998). More recently, large-scale pretraining on unlabelled corpora followed by supervised fine-tuning has become popular, effectively combining representation learning with task-specific supervision for NLP tasks (Devlin et al., 2019). This technique is particularly valuable in task-oriented dialogue modelling, where high-quality labelled data is scarce. It has been successfully applied both to the development of system modules (Heck et al., 2022; Peng et al., 2020) and to interactive dialogue environments (Lin et al., 2021).

## 2.9.2 Advantages and Limitations

Semi-supervised learning offers several benefits. It can dramatically improve performance when labelled data is scarce by exploiting unlabelled data, effectively enlarging the training set without incurring annotation costs. It also allows models to learn smoother and more generalizable decision boundaries, often improving robustness and reducing overfitting on small labelled datasets.

However, semi-supervised learning has its limitations. The quality of pseudo-labels or learned representations heavily influences performance; incorrect predictions can propagate errors and reinforce biases. Furthermore, semi-supervised methods often require careful hyperparameter tuning and task-specific design choices to balance the contributions of labelled and unlabelled data.

In summary, semi-supervised learning provides a practical compromise between fully supervised and unsupervised approaches, leveraging unlabelled data to enhance performance while retaining task-specific guidance. It is particularly relevant in task-oriented dialogues where annotated resources are limited but raw text is plentiful (He et al., 2022).

## 2.10 Reinforcement Learning

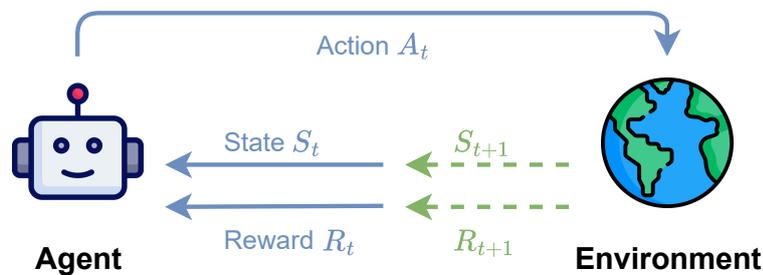


FIGURE 2.12: An illustration of the interaction between an RL agent and the environment. In each time step, the agent takes action  $A_t$  based on the observed state  $S_t$  from the environment. The agent then receives the reward  $R_{t+1}$  and the next state  $S_{t+1}$  from the environment.

The paradigms of supervised, unsupervised, and transfer learning all share a common reliance on static datasets: models are trained on collections of examples curated in advance, whether annotated or unlabelled. Reinforcement learning (RL), by contrast, shifts the focus from learning *from data* to learning *from experience*. In RL, an agent interacts directly with an environment, exploring sequences of actions and learning from feedback in the form of rewards or penalties. This interactive process allows the agent to optimise behaviour over time, not only to achieve immediate accuracy but also to maximise long-term success and user satisfaction.

### 2.10.1 Formal Definition

Reinforcement learning problems are typically formulated as a Markov Decision Process (MDP), defined by the tuple

$$M = (\mathcal{S}, \mathcal{A}, r, p, p_0), \quad (2.86)$$

where

- $\mathcal{S}$  is the state space: the set of states that the agent can visit,
- $\mathcal{A}$  is the action space: the set of actions that the agent can take,
- $r$  is the reward function, which maps the tuple  $(s, a)$  to a real value  $r(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ ,

- $p(s' | s, a)$  is the transition probability function and models the probability of transitioning into new state  $s'$  from state  $s$  after taking action  $a$ ,
- $p_0(s)$  is the probability of the starting state  $s$ ,

At each time step  $t$ , the agent observes a state  $s_t \in \mathcal{S}$ , selects an action  $a_t \in \mathcal{A}$  according to a policy  $\pi(a | s)$ , transitions to a new state  $s_{t+1}$ , and receives a reward  $r_t = r(s_t, a_t)$ .

### 2.10.2 Return and Value Functions

The goal of reinforcement learning is to find an optimal policy  $\pi^*$  that maximises the expected discounted return. Formally, the discounted return is defined as

$$G_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}, \quad (2.87)$$

where  $\gamma \in [0, 1)$  is the discount factor weighting future rewards.

The state-value function quantifies the expected return when starting from state  $s$  and following policy  $\pi$ :

$$V^\pi(s) = \mathbb{E}_\pi [G_t | S_t = s]. \quad (2.88)$$

With the state-value function, we can define the optimality of a policy. A policy  $\pi_1$  is considered better than another policy  $\pi_2$  if

$$V^{\pi_1}(s) \geq V^{\pi_2}(s), \quad \forall s \in \mathcal{S} \quad (2.89)$$

Therefore, a policy  $\pi$  is considered to be optimal if  $V^\pi(s) = \max_{\pi} V^\pi(s) \quad \forall s \in \mathcal{S}$ .

The action-value function considers the expected return when taking action  $a$  in state  $s$ :

$$Q^\pi(s, a) = \mathbb{E}_\pi [G_t | S_t = s, A_t = a]. \quad (2.90)$$

The optimal state-value and action-value functions define the best possible performance in a Markov decision process and are defined as

$$V^*(s) = \max_{\pi} V^\pi(s), \quad Q^*(s, a) = \max_{\pi} Q^\pi(s, a) \quad (2.91)$$

The advantage function measures how much better in terms of expected return when taking action  $a$  in state  $s$ :

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \quad (2.92)$$

### 2.10.3 Estimation of Value Functions

The general strategy for approximating the value functions is to move the approximations closer towards a target:

$$V(S_t) \leftarrow V(S_t) + \alpha (v_{\text{target}} - V(S_t)). \quad (2.93)$$

There are different approaches for defining the target. Monte Carlo (MC) methods estimate value functions by averaging returns from complete episodes. Temporal Difference (TD) methods update value estimates based on bootstrapping, using current estimates of future values rather than waiting for full returns.

$$\text{MC-target: } v_{\text{target}} = G_t \quad (2.94)$$

$$\text{TD-target: } v_{\text{target}} = R_t + \gamma V(S_{t+1}) \quad (2.95)$$

TD methods can generalise further. N-step TD improves upon one-step TD by looking ahead  $n$  steps to update value estimates, and  $TD(\lambda)$  interpolates multiple  $n$ -step returns using a hyperparameter  $\lambda$  to balance bias and variance.

### 2.10.4 Function Estimation in Deep Reinforcement Learning

In the methods we have examined in Section 2.10.3, a table is maintained to store the current estimates of value  $V(s)$  (or  $Q(s, a)$ ) for each state  $s$  (or state-action pair  $(s, a)$ ). While the tabular approach provides theoretical convergence guarantees, it does not scale to large or continuous state and action spaces. Moreover, it lacks the ability to generalise between similar states or actions, as each entry is updated independently.

To overcome these limitations, deep reinforcement learning (deep RL) replaces the explicit table with a neural network function approximator. The value functions  $V_\psi$  and  $Q_\phi$  are approximated with neural networks parameterised with  $\psi$  and  $\phi$  respectively. The network parameters are updated by minimising the prediction error between the current estimate and the target value. The update rule for  $V_\psi(S_t)$  is given by

$$\psi \leftarrow \psi + \alpha (v_{\text{target}} - V_\psi(S_t)) \cdot \nabla_\psi V_\psi(S_t), \quad (2.96)$$

where  $\alpha > 0$  denotes the learning rate. This corresponds to solving a regression problem with a mean-squared error (MSE) loss:

$$\mathcal{L} = \frac{1}{2} (v_{\text{target}} - V_\psi(S_t))^2. \quad (2.97)$$

The gradient of this loss with respect to the network parameters provides the update direction. By employing neural networks for function approximation, deep RL achieves the ability to generalise across large and continuous domains, an essential property for modern reinforcement learning applications, including robotics, control, and interactive dialogue systems.

### 2.10.5 Policy Optimisation

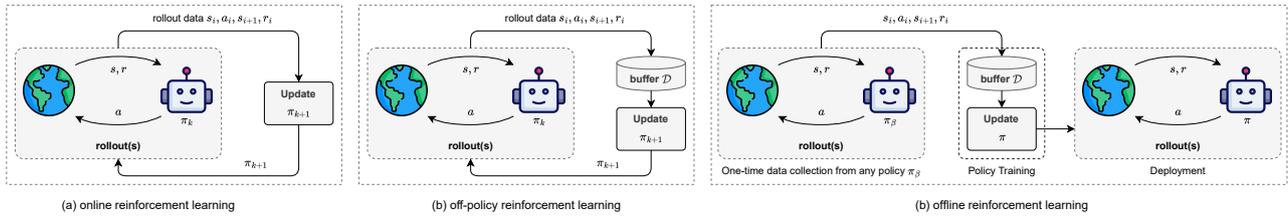


FIGURE 2.13: An illustration of (a) online reinforcement learning, (b) off-policy reinforcement learning, and (c) offline reinforcement learning.

### Value-based and Policy-based Approaches

RL algorithms can be broadly categorised into value-based and policy-based approaches. Value-based methods, such as Q-learning (Watkins and Dayan, 1992), aim to approximate the optimal action-value function  $Q^*(s, a)$  and derive the optimal policy by selecting actions that maximise Q-values:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[ r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right]. \quad (2.98)$$

Policy-based methods instead directly parameterise the policy  $\pi_\theta(a | s)$  and optimise it to maximise expected return. The policy gradient theorem (Sutton et al., 1999) provides the foundation:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(a | s) Q^\pi(s, a)]. \quad (2.99)$$

These form the basis of REINFORCE (Williams, 1992), Proximal Policy Optimisation (PPO) (Schulman et al., 2017), and actor-critic algorithms (Konda and Tsitsiklis, 1999), which combine value estimation with policy optimisation to stabilise training.

### Online and Offline Reinforcement Learning

Another important distinction among RL approaches is between *online* and *offline* RL. In online RL, the agent interacts with the environment during training, continuously collecting new trajectories to update its policy. This setting is flexible but can be expensive or unsafe when interactions are costly, as in human-in-the-loop dialogue. Offline RL, also known as *batch RL*, instead relies on a fixed dataset of past trajectories, without further interaction with the environment. While safer and more practical in scenarios where exploration is constrained, offline RL poses unique challenges, such as distributional shift between the training dataset and the policies being learned.

### On-policy and Off-policy Approaches

Reinforcement learning algorithms are also classified as *on-policy* or *off-policy*. On-policy algorithms, such as SARSA and policy gradient methods, evaluate and improve the same policy that is used to generate data. This typically yields stable updates but requires continuous exploration and data collection. Off-policy algorithms, by contrast, learn about a target policy while reusing data generated by a different behaviour policy. Q-learning and Deep Q-Networks (DQN) are classic examples of off-policy methods, enabling greater sample efficiency by reusing past experience. In dialogue systems, off-policy learning is particularly appealing, since training can proceed from logged conversations without requiring continual interaction with users.

#### 2.10.6 Advantages and Limitations

RL offers several advantages for natural language processing applications. By optimising behaviour through interaction, RL allows models to move beyond local accuracy and instead target long-term objectives such as sustained user satisfaction and task success. This makes RL particularly appealing for dialogue systems, where sequential decision-making and adaptability are essential. The ability to incorporate feedback—whether from users, reward models, or evaluation metrics—also provides a flexible framework for aligning system behaviour with desired outcomes.

Despite its promise, applying RL in NLP presents several challenges. Reward design is notoriously difficult, especially for subjective goals such as informativeness, coherence, or politeness. Sample inefficiency is another obstacle, as many RL algorithms require extensive exploration, which is impractical in human-in-the-loop settings. Finally, policy optimisation methods can be unstable without careful regularisation, making robust deployment difficult.

Nevertheless, RL provides a powerful framework for interactive and adaptive NLP systems, but its success depends on careful handling of reward specification, training efficiency, and stability.

## 2.11 Large Language Models

The scaling of unsupervised pretraining has given rise to large language models (LLMs), which represent a transformative advance in NLP. LLMs are trained on massive corpora with billions of parameters, typically in Transformer architectures (Vaswani et al., 2017). Through scale, LLMs achieve remarkable generalisation, enabling zero-shot, few-shot, and in-context learning without task-specific supervision.

### 2.11.1 Scaling Laws

Empirical studies by Kaplan et al. (2020) demonstrated that performance improves predictably as a power-law function of model size, dataset size, and compute budget. Formally, for loss  $\mathcal{L}$ , model parameters  $N$ , dataset size  $D$ , and compute  $C$ , the expected loss follows:

$$\mathcal{L}(N, D, C) \approx \mathcal{L}_\infty + aN^{-\alpha} + bD^{-\beta} + cC^{-\gamma},$$

where  $\mathcal{L}_\infty$  is the irreducible loss and  $(\alpha, \beta, \gamma)$  are scaling exponents estimated empirically. These results have motivated the training of ever-larger models such as GPT-3 (175B parameters; Brown et al., 2020) and PaLM (540B parameters; Chowdhery et al., 2023).

### 2.11.2 Training and Alignment

LLMs are typically pretrained on large corpora with autoregressive language modelling objectives, as we have discussed in Section 2.8.1.

While pretrained LLMs capture linguistic and factual knowledge, their outputs are not always aligned with human preferences. **Reinforcement learning from human feedback (RLHF)** has become a standard method for aligning models (Christiano et al., 2017; Ouyang et al., 2022).

RLHF frames language modelling as a reinforcement learning problem. Given an input prompt  $\mathbf{x}$ , the model generates an output  $\mathbf{y} \sim \pi_\theta(\mathbf{y} | \mathbf{x})$ , where  $\pi_\theta$  is the current policy (the LLM). Human annotators provide comparisons between outputs, which are used to train a reward model  $R_\phi(\mathbf{x}, \mathbf{y})$  that predicts preference scores. The LLM is then optimised to maximise the expected reward under this learned signal.

Formally, the RLHF pipeline consists of three stages.

**(i) Supervised fine-tuning (SFT).** A pretrained model is first fine-tuned on a small dataset of human demonstrations to initialise a policy  $\pi_{\text{SFT}}$  that produces higher-quality responses. This provides a strong starting point for preference-based optimisation.

**(ii) Reward modelling.** Human annotators rank model outputs  $\{y^{(1)}, \dots, y^{(k)}\}$  for a given prompt  $x$ . These comparisons are used to train a reward model  $R_\phi(x, y)$ , typically with a Bradley-Terry loss (Bradley and Terry, 1952):

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x, \mathbf{y}^{(i)}, \mathbf{y}^{(j)})} \left[ \log \sigma(R_\phi(x, \mathbf{y}^{(i)}) - R_\phi(x, \mathbf{y}^{(j)})) \right], \quad (2.100)$$

where  $\sigma$  is the logistic function, and  $\mathbf{y}^{(i)}$  is preferred to  $\mathbf{y}^{(j)}$  in the human ranking.

**(iii) Reinforcement learning.** Using the reward model as a surrogate for human preferences, the policy is further fine-tuned with reinforcement learning. The optimisation objective is to maximise expected reward:

$$J(\pi_\theta) = \mathbb{E}_{\mathbf{y} \sim \pi_\theta(\cdot | \mathbf{x})} [R_\phi(\mathbf{x}, \mathbf{y})]. \quad (2.101)$$

The policy gradient theorem gives the update rule:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(\mathbf{y} | \mathbf{x}) R_\phi(\mathbf{x}, \mathbf{y})]. \quad (2.102)$$

In practice, Proximal Policy Optimisation (PPO) (Schulman et al., 2017) is widely used to stabilise training by constraining the policy update between iterations.

Through this formulation, RLHF reframes alignment as policy optimisation under a learned reward function that encodes subjective human preferences. This connects maximum likelihood pretraining (optimising for next-token prediction) with reinforcement learning, where the objective is to generate helpful, harmless, and honest outputs in line with human expectations.

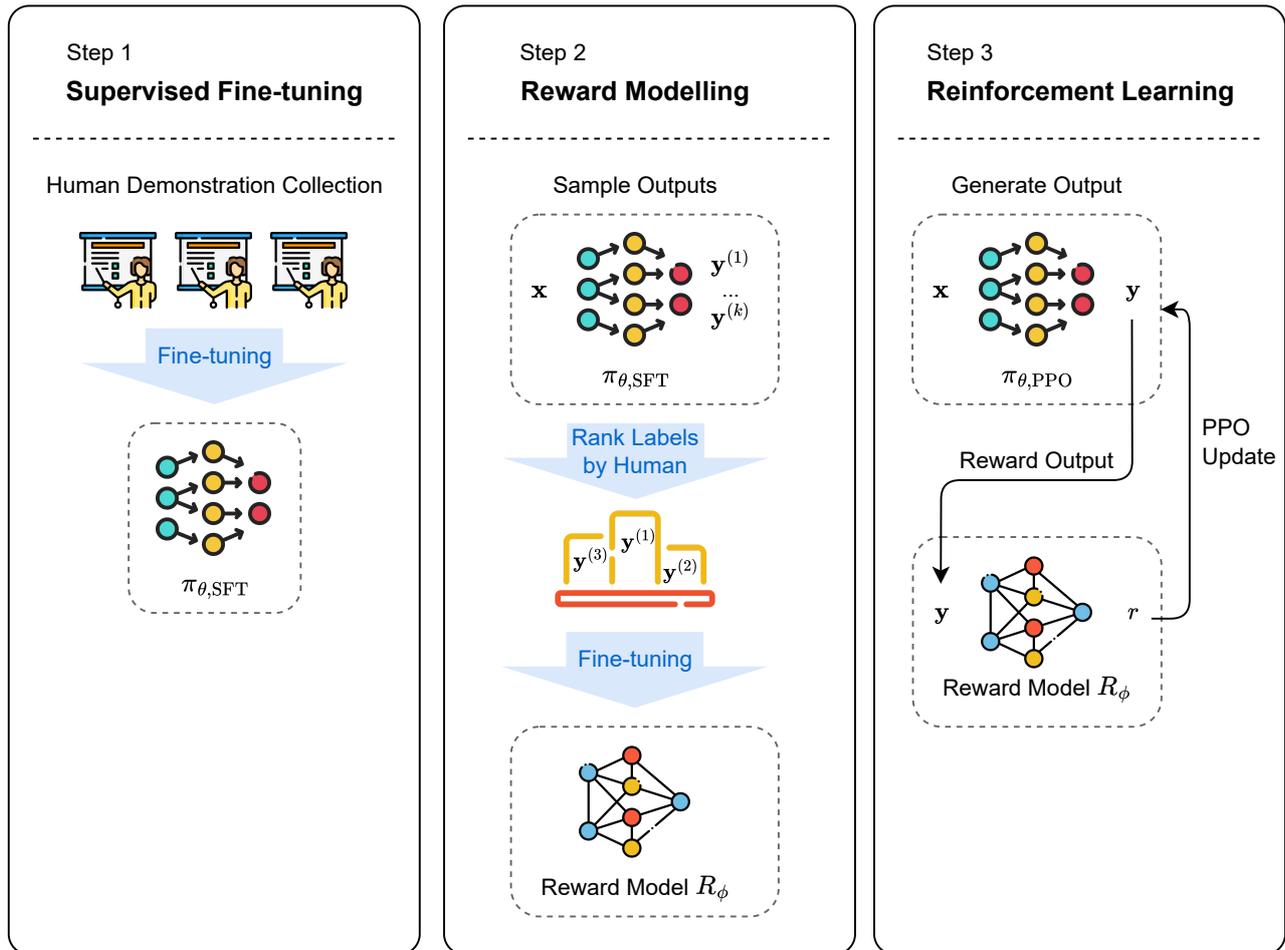


FIGURE 2.14: An illustration of Reinforcement Learning from Human Feedback (RLHF) for training a large language model (LLM). It follows three steps: (1) human demonstrations are collected and they are used to initialise the policy  $\pi_{\theta, \text{SFT}}$  (the LLM) via supervised fine-tuning (SFT) using the autoregressive language modelling loss in Equation 2.83. (2) For each prompt  $x$ , outputs  $y^{(1)}, \dots, y^{(k)}$  are sampled from the policy. Human annotators rank these sampled outputs to create a preference dataset, which is used to train the reward model  $R_\phi$ , following the loss in Equation 2.100. (3) For each new prompt  $x$ , the policy generates an output  $y$ . The reward model assigns a reward  $r$  for reinforcement learning of the policy following the gradient in Equation 2.102.

**Task-specific Fine-tuning** After pretraining, fine-tuning paradigms adapt LLMs for downstream tasks. Classical fine-tuning updates all parameters on task-specific data, but this is often computationally expensive. Low-rank adaptation (see Section 2.7.4) is a widely adopted approach to reduce the number of trainable parameters while retaining strong performance.

### 2.11.3 Emergent Abilities

A striking property of LLMs is the emergence of capabilities not present in smaller models (Wei et al., 2022b). We discuss some iconic abilities in this section.

**Zero-shot generalisation** LLMs can perform tasks with no labelled examples, relying only on task descriptions in natural language. This reflects the capacity of pretraining to endow models with broad functional knowledge.

**Few-shot learning** Given a prompt containing  $k$  labelled examples  $\{(x_i, y_i)\}_{i=1}^k$ , the model conditions on these examples and predicts the label for a new query  $x_{k+1}$ . This corresponds to conditional inference:

$$p(y_{k+1} \mid x_{k+1}, \{(x_i, y_i)\}_{i=1}^k). \quad (2.103)$$

**In-context learning** More generally, LLMs exhibit the ability to treat input text as a form of supervision, conditioning directly on task descriptions, instructions, or examples without updating model parameters. This blurs the line between supervised and unsupervised paradigms, as adaptation occurs purely through conditioning.

**Reasoning and Compositionality** LLMs also exhibit emerging reasoning capabilities. Chain-of-thought prompting (Wei et al., 2022a) encourages models to generate intermediate reasoning steps, which improves performance on arithmetic, logic, and commonsense tasks. Formally, instead of directly predicting the final answer  $y$ , the model is conditioned to produce a sequence of reasoning steps  $z(z_1, \dots, z_m)$ :

$$p(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) p(\mathbf{z} \mid \mathbf{x}). \quad (2.104)$$

This decomposition allows reasoning to be represented as latent structures, improving interpretability and problem-solving ability. Other methods, such as self-consistency (Wang et al., 2023), sample multiple reasoning chains and aggregate their outputs to improve robustness.

#### 2.11.4 Advantages and Limitations

Large language models bring several advantages. Their scale enables emergent abilities such as in-context learning and zero-shot transfer, dramatically reducing the need for task-specific training data. They serve as flexible general-purpose learners, achieving strong performance across diverse NLP benchmarks. Techniques such as parameter-efficient fine-tuning and RLHF further extend their adaptability, making it possible to deploy LLMs in practical and specialised applications (Sun et al., 2024).

At the same time, LLMs face significant limitations. Training and deploying models at scale requires enormous computational and energy resources, raising concerns about efficiency and environmental cost. Their outputs may include hallucinations (Huang et al., 2025), i.e. fluent but factually incorrect statements, and they can inherit or amplify biases present in their training data (Gallegos et al., 2024). Safety also remains a critical challenge as LLMs can generate harmful, offensive, or misleading content if prompted inappropriately (Sun et al., 2025). Finally, alignment with human values is still an open problem, as current methods such as RLHF rely on imperfect reward models and labour-intensive human feedback.

In summary, LLMs represent a paradigm shift in NLP, combining scale, emergent abilities, and flexible adaptation. Yet realising their full potential requires addressing critical challenges in efficiency, factuality, bias, and alignment.

## 2.12 Conclusion

This chapter surveys data-driven approaches in modern machine learning, which are fundamental to the work in this thesis. We also discuss the principal paradigms through which machine learning models acquire knowledge from data and experience. We began with supervised learning, the most established framework, where models learn from annotated input-output pairs. While powerful and conceptually clear, supervised learning is fundamentally constrained by the availability of labelled data. Transfer learning addresses this limitation by leveraging knowledge acquired in one domain to improve performance in another, with techniques such as domain adaptation, data augmentation, multi-task learning, and parameter-efficient fine-tuning proving especially effective in NLP.

We then turned to unsupervised learning, which exploits the abundance of raw text to learn representations without explicit supervision. Objectives such as language modelling, autoencoding, and contrastive learning have become foundational for modern NLP, enabling large-scale pretraining. Reinforcement learning, by contrast, shifts the paradigm from learning from static data to learning from interaction and experience. In NLP, RL provides a framework for optimising long-term objectives in interactive systems such as dialogue agents, though challenges remain in reward specification, sample efficiency, and stability.

Finally, we examined large language models (LLMs), which synthesise these paradigms at an unprecedented scale. Through massive unsupervised pretraining, scaling laws, and emergent abilities such as in-context learning, LLMs have become general-purpose learners. Fine-tuning strategies, parameter-efficient adaptation, and reinforcement learning from human feedback extend their applicability and align their behaviour with human preferences, while reasoning techniques such as chain-of-thought prompting push the limits of their compositional abilities. At the same time, limitations in efficiency, factual reliability, and bias highlight the need for continued innovation.

In summary, the trajectory from supervised to transfer, unsupervised, reinforcement, and large-scale learning reflects a broadening of perspective: from narrow, task-specific training towards flexible, adaptive systems capable of leveraging vast amounts of data and experience. These paradigms collectively provide the methodological foundation upon which this thesis builds, and the next chapter will situate them in the specific context of dialogue modelling.

## Chapter 3

# Affective Computing

### 3.1 Overview

The scientific study of affect and emotion has a long history across psychology, philosophy, and neuroscience, with successive theories offering different perspectives on how emotions arise. Darwin's evolutionary view emphasised that emotions and their expressions are biologically hardwired, universal across cultures, and adaptive for survival (Darwin, 1872). Building on the link between body and feeling, the James-Lange theory proposed that emotions are the result of perceiving physiological changes in the body, such that we feel afraid because we notice ourselves trembling (James, 1884; Lange, 1885). In contrast, the Cannon-Bard theory challenged this sequence, arguing that emotions and bodily responses occur simultaneously as parallel outcomes of brain processing, rather than one causing the other (Cannon, 1927). Later, with the rise of cognitive psychology, appraisal theories shifted attention to the role of interpretation and meaning-making, suggesting that emotions emerge from individuals' evaluations of situations in relation to their goals and contexts (Lazarus, 1991; Scherer, 2005). These perspectives trace a progression from viewing emotions as innate biological reactions, to bodily feedback, to brain-based processes, and finally to cognitively mediated evaluations. These theories lay foundations that continually shape computational models linking affect with cognition, physiology, and behaviour.

These theoretical perspectives motivated the development of different ways to model and classify emotions in both psychology and computational research. The emphasis on universality and biological hardwiring in Darwin's view inspired categorical approaches, such as the theory of basic emotions by Ekman (1971), where emotions are treated as discrete and fundamental units. While these categorical models have been influential, especially in early affect recognition, they have also been criticised for oversimplifying the diversity and context-dependence of emotional experience, as well as for assuming cross-cultural universality that has been increasingly questioned (Russell, 1994). The James-Lange and Cannon-Bard theories highlighted the importance of both physiological responses and central brain processes in shaping emotion. These paved the way for moving beyond discrete categories and inspiring dimensional models such as the Circumplex framework (Russell, 1980), which describe emotions along continuous axes like valence and arousal. However, dimensional models abstract away from the cognitive and social factors that shape emotional meaning. These theories focusing on internal mechanisms offer limited guidance for modelling the dynamic, multi-modal, and contextually embedded nature of emotion observed in the real world. Finally, cognitive appraisal theories, with their focus on evaluation and context, informed more complex frameworks such as the Ortony, Clore, and Collins (OCC) model (Ortony et al., 1988), which represent emotions as dynamic processes shaped by goals, meaning, and situations. Despite their conceptual richness, appraisal-based models can be challenging to formalise and operationalise computationally due to the complexity and subjectivity of the appraisal processes they describe. Despite their respective strengths and limitations, these taxonomies and models of emotion offer complementary perspectives for representing affective states in computational systems, and they collectively inform the conceptual foundations underlying the design of EmoWOZ (Chapter 5).

Building on foundational theories of emotion and recognising the central role of affect in communication (Frijda, 1986), decision-making (Damasio, 1994), and social bonding (Ekman, 1992a;

Parkinson, 1996), the field of *affective computing* was formally introduced (Picard, 1997). It focuses on designing systems that can automatically recognise, interpret, and respond to human emotions, enabling interactions that are not only functional but also socially and emotionally aware. This concept parallels Mayer and Salovey's notion of emotional intelligence in human beings (Salovey and Mayer, 1990), defined as the ability to perceive, understand, and use emotions effectively in oneself and others. Just as emotional intelligence guides humans in adapting behaviour to social and emotional contexts, affective computing seeks to embed similar capabilities in machines, enabling machines to respond to users in ways that are emotionally attuned and contextually appropriate.

The applications of affective computing machines are diverse and expanding. In open-domain chat applications, empathetic chatbots can reduce frustration and improve satisfaction (Prendinger and Ishizuka, 2005). In healthcare, affect-aware systems can support mental health monitoring, stress detection, and patient-clinician communication (Luneski et al., 2008; Schmidt et al., 2019). In education, intelligent tutoring systems leverage affect recognition to adapt feedback and maintain learner engagement (D'mello and Graesser, 2013). In entertainment and gaming, affective models enhance immersion and personalisation (Yannakakis and Hallam, 2007). Most importantly for this thesis, in task-oriented dialogue systems, we argue that affective computing enables systems to detect, process, and express emotions during goal-directed interactions, ensuring smoother collaboration between human users and artificial intelligence (AI) systems for the ultimate task success.

In summary, affective computing is grounded in studies of the origins of emotion and their modelling. It represents a convergence of theories and methods from multiple disciplines, aiming to endow machines with the ability to sense, interpret, and act upon human emotions, akin to having emotional intelligence. The field provides the conceptual and technical foundations for emotionally intelligent task-oriented dialogue systems, which are crucial to the work presented in this thesis and will be explained in greater detail in this chapter.

## 3.2 Taxonomy of Emotion

Understanding how to represent emotions is fundamental not only for psychology but also for affective computing. Over the years, researchers have proposed various taxonomies of emotion that aim to capture the diversity, structure, and dynamics of affective states. These taxonomies provide formal frameworks for describing emotions in ways that can be consistently studied, compared, and utilised within computational systems. Emotions can be defined with dimensional models, which view emotions as points in a continuous affective space, and categorical models, which treat emotions as discrete, fundamental units. Each approach reflects different theoretical assumptions about the nature of emotion and offers unique utilities for affective computing.

### 3.2.1 Dimensional Models of Emotion

Dimensional models conceptualise emotions as points in a continuous affective space, rather than as discrete categories. These models attempt to identify the latent dimensions that underlie diverse emotional experiences. As early as 1897, Wilhelm Max Wundt, often regarded as the father of modern psychology, proposed to describe emotions along three axes: "pleasurable versus unpleasurable", "arousing versus subduing", and "strain versus relaxation" (Høffding, 1892). In 1954, Harold Schlosberg identified three related dimensions of emotion: "pleasantness-unpleasantness", "attention-rejection", and "level of activation" (Schlosberg, 1954). These early conceptualisations of emotion dimensions laid the foundation for subsequent formal frameworks that continue to shape contemporary affective computing.

#### Two-dimensional Models

**The Circumplex Model of Affect** One of the earliest and most influential frameworks is Russell's Circumplex Model of Affect (Russell, 1980). It organises affective states in a circular structure defined

by two orthogonal axes: valence (pleasant-unpleasant) and arousal (low-high activation). Emotions are represented as coordinates within this space, such as joy (high valence and high arousal) or sadness (low valence and low arousal). The circumplex provides an account of affective experience and aligns with evidence from psychophysiology that valence and arousal map onto distinct neural and autonomic systems. This taxonomy forms the foundation for many other dimensional models discussed in this section and for hybrid approaches described in Section 3.2.3.

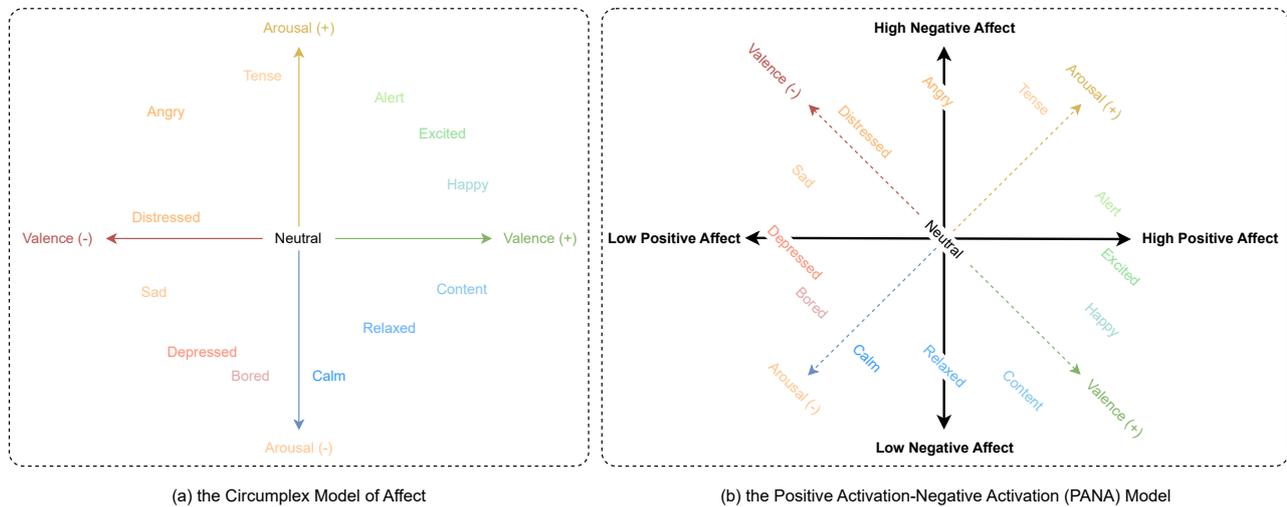


FIGURE 3.1: An illustration of (a) the Circumplex Model of Affect, and (b) the positive activation-negative activation (PANA) model. Conceptually, PANA can be viewed as a 45-degree rotation of the Circumplex Model of Affect.

**The Positive Activation-Negative Activation Model** An alternative but related framework is the Positive Activation-Negative Activation (PANA) model (Watson and Tellegen, 1985). The PANA model defines affective space using two broad dimensions: positive activation (engagement, enthusiasm) and negative activation (distress, fear). Conceptually, PANA can be viewed as a 45-degree rotation of the Circumplex Model of Affect, as illustrated in Figure 3.1. Unlike valence-arousal models, which are bipolar along each axis, PANA treats positive and negative activation as partially independent, allowing for states such as “mixed emotions”, which have simultaneous high positive and high negative activation. For example, consider a person about to give an important public speech: they may feel excitement, enthusiasm, and energy about the opportunity (high positive activation), while simultaneously experiencing anxiety, nervousness, or fear of failure (high negative activation). The PANA model offers an alternative perspective on affective space that highlights the independence of positive and negative activation dimensions.

**The Vector Model** The Vector Model of emotion (Bradley et al., 1992) builds on the observation that positive and negative affect often vary in opposite directions. As shown in Figure 3.2, it assumes that affective states are organised along two vectors radiating from a common neutral point: one vector representing increasing positive affect and the other representing increasing negative affect. Rather than treating positive and negative affect as fully independent dimensions, as in the PANA model, the Vector Model assumes a partially inverse relationship, such that strongly positive and strongly negative emotions are unlikely to co-occur. This framework captures the idea that emotional experiences tend to cluster toward either end of the affective spectrum, with ambivalence or neutrality represented near the origin. The Vector Model has been influential in the study of mood and well-being, particularly in psychology and health sciences, though its assumption of mutual exclusivity has been debated in contexts such as conversational and social interaction data, where mixed emotions are common.

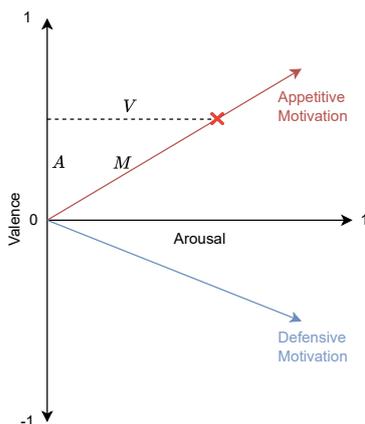


FIGURE 3.2: An illustration of the Vector Model of emotion.

Among these models, the valence dimension, in particular, has become the most widely used, inspiring sentiment analysis tasks such as SemEval 2007 Task 14 (Strapparava and Mihalcea, 2007), SemEval-2017 Task 4 (Rosenthal et al., 2017), and datasets like the Movie Review Dataset (Pang and Lee, 2005) and Stanford Sentiment Treebank (Socher et al., 2013). Beyond text, valence and dominance labels are also applied in other modalities, including images (Dan-Glauser and Scherer, 2011) and human-computer interaction studies (Broekens and Brinkman, 2013), providing a cross-domain framework for representing affective responses.

### Three-dimension Models

**The Pleasure-Arousal-Dominance Model** Building on the Circumplex Model of Affect, Mehrabian and Russell introduced the Pleasure-Arousal-Dominance (PAD) model in the context of environmental psychology (Mehrabian and Russell, 1974). Here, pleasure closely corresponds to valence, arousal captures intensity, and dominance reflects the degree of perceived control in a situation. PAD has been widely applied to describe affective reactions to environments (Robert and John, 1982), and interpersonal interactions (Mehrabian, 1996).

**The Valence-Arousal-Dominance Model** The PAD model was later reformulated in computational contexts as the VAD (Valence-Arousal-Dominance) model (Figure 3.3). Bradley and Lang’s Affective Norms for English Words (ANEW) dataset (Bradley and Lang, 1999) popularised this terminology, providing large-scale ratings of words on the VAD dimensions. Subsequent expansions, such as the work by (Warriner et al., 2013), made VAD the de facto standard in natural language processing (NLP) and affective computing, where continuous, scalable representations are crucial. Representative datasets such as EmoBank (Buechel and Hahn, 2017) further extend this framework by providing VAD annotations for a wide range of sentences, enabling large-scale modelling of emotional content in written language and supporting advanced applications in sentiment analysis, emotion recognition, and computational affective modelling.

### Strengths and Limitations of Dimensional Models

From the perspective of affective computing, dimensional models offer several advantages. They provide continuous, fine-grained representations of affect that are amenable to statistical learning and time-series modelling. They facilitate the capture of gradual changes and trajectories of emotion over time. Furthermore, they align with the fact that physiological and linguistic signals often vary continuously rather than discretely.

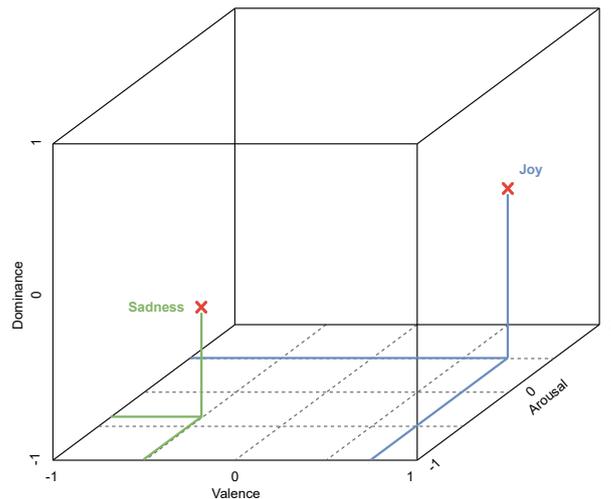


FIGURE 3.3: An illustration of the Valence-Arousal-Dominance model of emotions.

At the same time, their abstraction may obscure the qualitative distinctiveness of specific emotions (e.g., anger vs. fear), motivating the complementary use of categorical and hybrid frameworks (see Sections 3.2.2 and 3.2.3).

### 3.2.2 Categorical Models of Emotion

In contrast to dimensional models, categorical models conceptualise emotions as distinct, qualitatively different states rather than points in a continuous space. Each emotion is treated as a fundamental unit of experience, often associated with characteristic facial expressions, physiological patterns, and behaviours. Categorical models have been particularly influential in affective computing, providing clear labels that can be operationalised for recognition and synthesis in computational systems.

#### The Theory of Basic Emotions

The work of Ekman (1971) on basic emotions proposed six culturally universal and biologically hardwired emotions: happiness, sadness, anger, fear, disgust, and surprise (also known as the “Big Six” theory). These emotions are associated with distinct facial expressions, physiological responses, and adaptive functions, which allow efficient communication and social coordination. Ekman’s framework has been foundational in affective computing, particularly in emotion recognition from facial (Goodfellow et al., 2013) and vocal cues (Cao et al., 2014; Pichora-Fuller and Dupuis, 2020), providing a set of clear, discrete labels that can be reliably annotated and operationalised in computational systems (Velásquez, 1997). While this set of basic emotions has been widely adopted, it is often adapted to specific domains for more accurate description, as seen in datasets such as the Berlin Emotional Speech Database (Burkhardt et al., 2005), or expanded to incorporate more nuanced affective states, as in IEMOCAP (Busso et al., 2008). These modifications highlight both the strengths and limitations of the “Big Six” framework, and underscore the need for broader and more flexible taxonomies of emotion.

#### Expanded Taxonomies of Emotions

**Differential Emotions Theory** Building on Ekman’s foundation, Izard (1977) introduced a more nuanced set of ten fundamental emotions: interest, joy, surprise, sadness, anger, disgust, contempt, fear, shame, and guilt. This taxonomy of emotions emphasises the motivational and cognitive roles of emotions in shaping attention, appraisal, and learning beyond mere facial universals.

**Cognitive-Motivational-Relational Theory of Emotion** Lazarus (1991) proposed a framework of fifteen emotions, grounded in appraisal theory, where emotions arise from individuals' evaluations of events relative to their goals, values, and coping potential. The list includes emotions such as anger, guilt, shame, pride, love, and envy, among others, and emphasises the relational and evaluative nature of affective experience. This framework has influenced affective computing by providing a more context-sensitive approach to emotion recognition, where situational appraisal informs the presence and intensity of discrete emotional states (Gratch and Marsella, 2004; Marsella et al., 2010).

### Data-Driven Emotion Taxonomies

With the recent advances in data-driven machine learning, there is a shift from emotion frameworks rooted in psychological theory to data-driven taxonomies, leveraging large-scale human perception studies and computational methods to identify patterns in emotional expression. For example, Cowen and Keltner (2017) identified 27 distinct emotions using large-scale studies of human labelling of facial expressions, vocalisations, and bodily gestures. This taxonomy includes nuanced affective states such as awe, nostalgia, embarrassment, desire, and relief, reflecting the rich diversity of human emotion beyond traditional basic categories. Data-driven approaches are also evident in social media analysis, where emojis and text reactions are used to derive emotion labels in large-scale datasets, capturing naturally occurring emotional expressions in context (Felbo et al., 2017). These methodologies complement traditional taxonomies by uncovering subtle, mixed, or context-dependent emotional states that are often underrepresented in theory-driven models.



FIGURE 3.4: An illustration of the expansion of categorical taxonomies: from emphasizing universality, to considering context and appraisal, and to observing from data.

### Strengths and Limitations of Categorical Models

Categorical models of emotion have evolved considerably over time, reflecting both theoretical insights and empirical findings. Figure 3.4 illustrates that the progression of emotion taxonomies follows a clear trend: as researchers incorporate richer theoretical perspectives and empirical evidence, the granularity and complexity of emotion taxonomies increase, allowing for more nuanced representation and modelling of human affect.

Categorical taxonomies offer several significant advantages for affective computing. They are intuitive and interpretable, making them easy to label and communicate across researchers and systems. They have a strong grounding in observable behaviour, such as facial expressions, vocal cues, or appraisal processes, which allows reliable annotation and operationalisation in computational models. Furthermore, categorical labels facilitate classification and retrieval in natural language processing and multimodal datasets, providing clear and consistent targets for supervised learning and evaluation.

At the same time, categorical models introduce notable challenges. They can oversimplify or overcomplicate the complexity of affective experiences, depending on the granularity of labels. Mixed or co-occurring emotions are difficult to represent accurately within strictly discrete categories, limiting the ability to capture nuanced affective states. Additionally, emotion categories can vary

across cultures and contexts, reducing the universality of any single taxonomy and complicating cross-domain or cross-cultural applications.

### 3.2.3 Hybrid Models of Emotion

Hybrid models integrate categorical and dimensional representations to provide more flexible and nuanced descriptions of affective states. They aim to preserve the interpretability of discrete categories while modelling gradations, intensities, and relational structure that are well captured in dimensional frameworks.

#### Plutchik's Wheel of Emotions

Plutchik's emotion model (Plutchik, 1980) is often considered a hybrid because it combines categorical and dimensional elements. As illustrated in Figure 3.5, the model identifies eight primary emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) as discrete categories while arranging them in a wheel that illustrates opposites, combinations, and varying intensities. For example, intensity is encoded in concentric layers, showing how emotions can range from mild to extreme (e.g., serenity  $\rightarrow$  joy  $\rightarrow$  ecstasy). Plutchik also defines secondary emotions arising from combinations of primary emotions (e.g., optimism = joy + anticipation), providing a structure that links discrete categories to continuous relationships. Plutchik's model has been used to annotate textual datasets for emotion detection (Mohammad et al., 2018). By providing both discrete categories and hierarchical relationships, Plutchik's model also facilitates more nuanced and interpretable affective computing applications (Qi et al., 2019). However, the shortcomings of categorical models still exist for Plutchik's taxonomy.

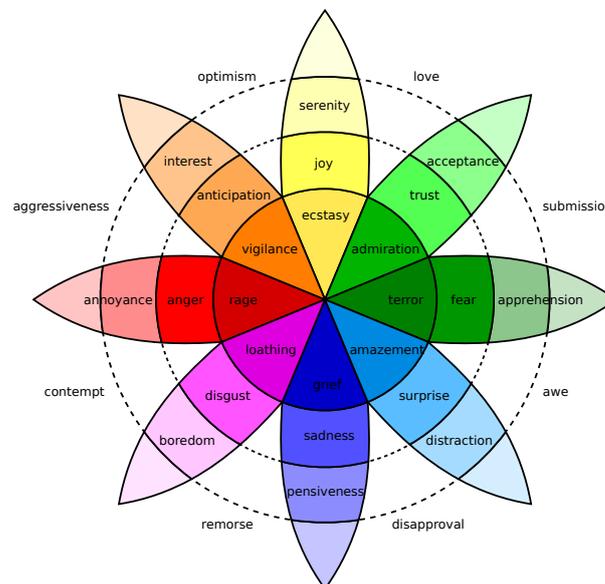


FIGURE 3.5: An illustration of Plutchik's wheel of emotions (Source: Wikimedia Commons, "Plutchik-wheel.svg"; License: Public Domain).

#### The Ortony, Clore, and Collins Model

The OCC model is another influential hybrid framework, based on cognitive appraisal theory (Ortony et al., 1988). Unlike Plutchik's taxonomy, the OCC model defines a more fine-grained set of 22 basic

emotion categories, each associated with a structured cognitive appraisal process, and provides a way to quantify the valence and intensity of emotions within specific scenarios.

As illustrated in Figure 3.6, the OCC model identifies emotion types derived from appraisals of three classes of elicitors, including events, agents, and objects. Emotions toward events (e.g., joy, distress, hope, fear) depend on the perceived desirability of outcomes relative to one's goals. Emotions toward agents (e.g., pride, admiration, anger, gratitude) depend on appraisals of the perceived appropriateness of an agent's actions, while emotions toward objects (e.g., love, hate) reflect attitudes shaped by personal preferences. The model also integrates quantitative variables such as desirability, likelihood, effort, and realisation, which jointly determine an emotion's valence and intensity. Putting together, as an example, the emotion category "relief/fears-confirmed" is elicited by an event that is expected to cause a consequence for oneself. The valence and intensity are determined by the desirability and likelihood of the consequence.

This rule-based structure makes the OCC model particularly attractive for computational emotion modelling. Because it specifies a finite set of appraisal-based conditions, it can be formalised in algorithms and applied in domains such as affective virtual agent (Gratch and Marsella, 2004; Paiva et al., 2005) and tutoring systems (Conati and Maclaren, 2009). Its explicit mapping from appraisal structures to emotion categories enables transparent reasoning about why an emotion is generated, an advantage over purely statistical or black-box approaches.

However, the OCC model has also been critiqued for several limitations. It assumes a relatively static and rule-based appraisal process, which can oversimplify the dynamic, context-dependent, and socially mediated nature of emotion in real interaction. Moreover, while it covers a broad taxonomy of emotions, some categories (e.g., reproach, gratitude) depend heavily on moral and cultural assumptions that may not generalise across contexts. Its reliance on manually defined appraisal rules also poses challenges for scalability and adaptation to real-world, noisy data.

Despite these limitations, the OCC model explicitly incorporates the notion of an *Elicitors*, the event, object, or agent that triggers an emotional response, which is particularly relevant and important in ToDs. In ToDs, user emotions are typically centred on task performance, and the dialogue context is limited, with clear elicitors such as specific events, the speaker (user or agent), and objects of interest as offered by the system. This explicit mapping between emotions and their elicitors makes the OCC model well-suited for capturing the causal structure of user affect in ToD interactions.

In this thesis, the OCC model informs the design of the EmoWOZ annotation scheme, which adapts its appraisal structure to the task-oriented dialogue context. EmoWOZ refines the OCC taxonomy by focusing on emotions relevant to task-oriented dialogue interactions, specifically those elicited by the event that motivates the user to engage with the system and by the agent, the information-providing system. Implementation details are provided in Chapter 5.

### Strengths and Limitations of Hybrid Models

Hybrid models of emotion offer a pragmatic balance between categorical and dimensional approaches, combining the interpretability of discrete labels with the flexibility of continuous representations. This dual structure allows them to capture nuanced and interrelated affective states. Such models are particularly valuable in computational applications, including dialogue systems, virtual agents, and natural language processing, where both distinct emotion categories and continuous affective features are required. At the same time, hybrid models introduce challenges: their increased complexity can make annotation and modelling more difficult, and the mapping between continuous dimensions and discrete categories can be ambiguous or context-dependent.

## 3.3 Emotion Intelligence in Computers

Building upon the theory of multiple intelligences by Gardner (1983), the concept of emotional intelligence (EI) was formalised by (Salovey and Mayer, 1990) as "the ability to monitor one's own

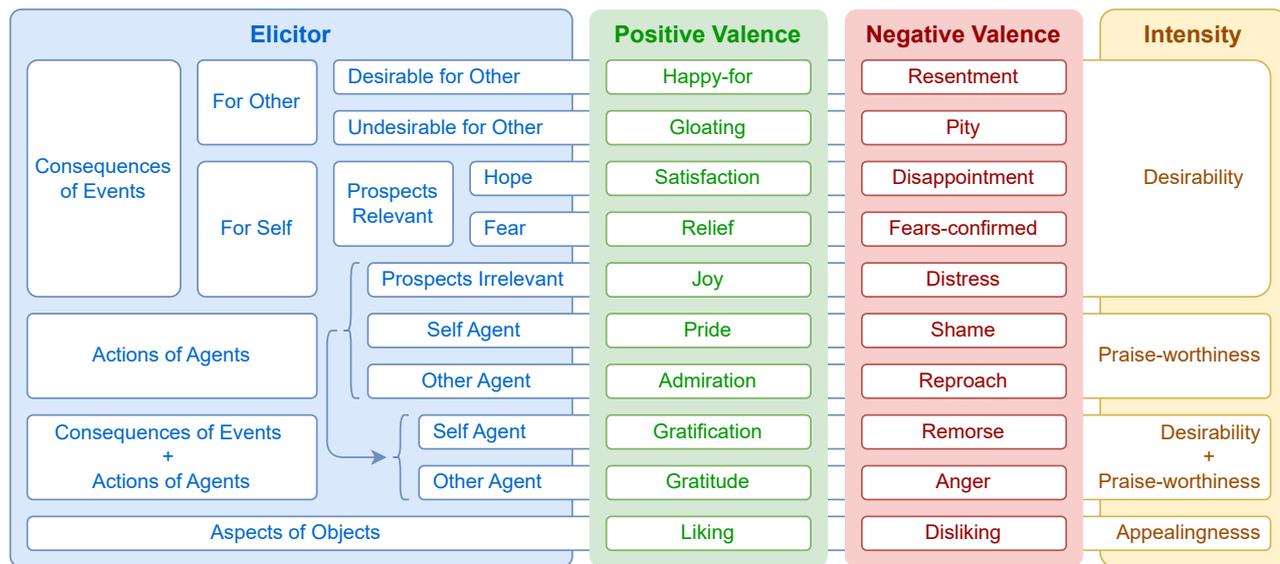


FIGURE 3.6: An illustration of the OCC model of emotions. In the OCC model, emotions are defined as valenced reactions to elicitors: consequences of events, actions of agents, and aspects of objects. The intensity depends on the attributes of the elicitor.

and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions." In its mature form (Mayer and Salovey, 1997), EI is described in four core skills:

1. **Emotion perception:** the ability to identify emotions in oneself and others through facial expressions, body language, and tone of voice;
2. **Emotion understanding:** the ability to analyse and differentiate complex emotion blends, to know the transitions between emotions, and to understand the causes and consequences of emotions;
3. **Emotion facilitation:** the ability to harness emotions to support reasoning and problem solving;
4. **Emotion management:** the ability to regulate one's own emotions and those of others effectively to achieve desired outcomes.

This framework, which represents a crucial cognitive and social competence for human beings, has been a central focus of study in affective computing, the interdisciplinary field concerned with equipping machines with similar capabilities.

Affective computing, pioneered by Picard (1997), maps these human emotional abilities to three core abilities of affective computing: to recognise, to express, and to "know" emotions. Emotion recognition allows the system to accurately identify a user's emotional state, be it a discrete emotion, a numerical value along a dimensional model, or a hybrid representation, and emotion expression enables the system to generate emotional responses that are readily recognisable and appropriate for the context. According to Picard, the third ability, to "know" emotions, carries a broader and more complex suite of abilities. This involves leveraging emotions to influence interaction dynamics and guide sequential decision-making for task success. In this sense, recognition and expression can be viewed as surface-level instantiations of such a deeper capacity: a system that "knows" emotions is able not only to represent them but also to strategically deploy them. To use a more formal term that also aligns with the semantic processing pipeline in task-oriented dialogues (ToDs) (Young, 2002) (as we introduce in greater detail in Chapter 4), we call this ability **emotion processing**.

In this section, we will discuss the approaches to developing each of these three core abilities of an affective computing system: to perceive, to express, and to process emotions.

### 3.3.1 Emotion Perception

The first core ability of an emotionally intelligent system is to **recognise emotions**, which corresponds to the ability of humans to perceive affective states in others. Within the field of affective computing, the computational analogue of emotion perception is most often framed as the task of emotion recognition or, more broadly, sentiment and affect analysis over textual, acoustic, or multimodal input.

Formally, let  $X = \{x_1, x_2, \dots, x_T\}$  denote a sequence of utterances in a dialogue, where each  $x_t$  represents the  $t$ -th utterance.

The task of *emotion classification* is to learn a mapping

$$f : x_t \mapsto y_t, \quad (3.1)$$

where  $y_t \in \mathcal{E}$  is an emotion label drawn from a predefined emotion set  $\mathcal{E}$  (categorical), or a continuous vector in  $\mathbb{R}^d$  (dimensional, in valence-arousal-dominance coordinates for example).

When extended to multi-turn dialogues, the problem is referred to as *emotion recognition in conversations* (ERC):

$$f : (x_1, \dots, x_t) \mapsto y_t, \quad (3.2)$$

where emotion recognition for each utterance  $x_t$  must consider both the conversational context, including preceding turns and speaker roles. This formulation captures phenomena such as repetition, coreference, and emotional trajectories. These phenomena are central to dialogue interactions and crucial to accurate emotion recognition.

ERC plays a crucial role in task-oriented dialogue systems. For instance, recognising that a user expresses *frustration* after a failed booking attempt enables the system to adapt its subsequent behaviour and maintain cooperation. However, challenges remain, including the scarcity of balanced conversational datasets, the prevalence of subtle or implicit emotion cues, and the need for real-time robustness in practical deployment.

### 3.3.2 Emotion Expression

The second ability is to **express emotions**, which mirrors the human ability to communicate affective states in socially recognisable ways. In natural language processing, this ability is instantiated in the task of natural language generation (NLG).

**NLG for open-domain Response Generation.** Unconstrained NLG, often employed in open-domain dialogue, formulates response generation as the problem of producing an utterance  $\hat{x}$  directly from dialogue context  $h_t = (x_1, \dots, x_t)$ :

$$\hat{x} \sim p_\theta(x | h_t). \quad (3.3)$$

Here,  $p_\theta$  is a parameterised language model that maximises the likelihood of next utterances given prior context. This paradigm enables fluent and contextually relevant responses. The emotion in the response is implicitly modelled based on the context. This method provides limited control over *what* is said and *how* it is said.

**Conditional natural language generation.** Conditional NLG can be extended to incorporate one or more attributes and has focused on two desiderata: *fidelity* to the semantic condition, and *fluency* of the generated utterance. With conditional NLG approaches, we can control the emotion conveyed in the response of a dialogue system:

$$\hat{x} \sim p_\theta(x | h_t, e), \quad (3.4)$$

where  $p_\theta$  is trained to model the probability distribution of tokens given a context  $h_t$  and the desired emotion  $e$ .

In ToDs, conditional NLG is used to produce responses that faithfully express the intended dialogue act semantics. In this thesis, we aim to extend the NLG of ToD systems by introducing an additional conditioning layer on emotion. While this enables the system to generate responses that are not only semantically accurate but also emotionally attuned, several challenges are presented:

- **Hallucination:** affective conditioning may induce inappropriate or factually unsupported content. For example, a system may fabricate special features of a restaurant while attempting to express positive emotions to engage the user.
- **Spurious correlations:** the system may pick up that apologies are frequently associated with the fact that nothing is found, so it may overlook situations where the system underperforms. During response generation, models may also rely on superficial cues (e.g., punctuation, lexical tokens) rather than genuine affective expression.
- **Balancing dual objectives:** ensuring semantic correctness while conveying the intended affect is particularly critical in task-oriented systems, where factual accuracy cannot be compromised.

In Chapters 8 and 9, we demonstrate how our proposed methods address these challenges and achieve a balance between semantic fidelity and emotional expressiveness.

### 3.3.3 Emotion Processing

The third and arguably most complex ability is to **process emotions**, corresponding to the human ability to facilitate decision-making and problem-solving using emotion. In affective computing, this ability is less explicitly formalised, often subsumed under the umbrella of empathetic response generation: the system's ability to understand a user's emotional state and respond in a manner that is contextually appropriate, supportive, or soothing.

However, empathetic response generation has remained largely data-driven, typically realised through supervised learning on conversational corpora that associate user utterances with emotionally appropriate responses. While such approaches can produce linguistically empathetic outputs, they offer little explicit control over the emotional intent of the response and lack reinforcement mechanisms to align generated empathy with conversational goals or long-term user satisfaction.

This limitation becomes even more pronounced in ToD systems, where emotional appropriateness must be balanced with task efficiency and factual accuracy. Unlike open-domain dialogue, ToD interactions require precise semantic control to ensure successful task completion (e.g., booking, information retrieval). Introducing emotion without a principled decision framework risks compromising task performance through inconsistent or overly affective responses. Therefore, a more structured formulation is needed to integrate emotion processing capability directly into the decision-making process rather than treating it as a post-hoc linguistic adaptation.

In Chapters 8 and 9, we demonstrate how we formally integrate emotion processing into dialogue policy learning and jointly optimise dialogue actions and emotional intents via reinforcement learning.

## 3.4 Conclusion

This chapter has reviewed the theoretical and computational foundations of affective computing, the interdisciplinary field concerned with endowing machines with the ability to sense, interpret, and act upon human emotions. We traced the intellectual roots of affective computing to classical theories of emotion in psychology and neuroscience, which progressed from evolutionary and physiological perspectives to appraisal-based and cognitive frameworks. These theories motivated diverse taxonomies of emotion, ranging from dimensional models that capture continuous affective variation to categorical models that emphasise discrete emotional states and hybrid approaches that integrate both. Such taxonomies provide the representational scaffolding required to operationalise emotions in computational systems.

Building on these foundations, affective computing has emerged as the computational analogue of emotional intelligence, mapping human capacities for perceiving, expressing, and processing emotions onto machine abilities. These abilities constitute the pillars of emotionally intelligent interaction. While affective computing has already demonstrated utility across domains such as healthcare, education, customer service, and entertainment, this thesis focuses on its application in ToDs. Being overlooked for long in this setting, emotionally intelligent behaviour is not a luxury but a necessity: user frustration, satisfaction, and trust directly influence whether task success can be achieved.

The three core tasks of affective computing align naturally with stages in the ToD pipeline. Accordingly, this chapter has discussed them in the context of ToDs: (1) *emotion perception*, where systems recognise users' affective states within dialogue contexts; (2) *emotion expression*, where systems generate semantically accurate yet emotionally conditioned responses; and (3) *emotion processing*, where dialogue policies balance task success with emotional appropriateness by selecting both functional actions and affective strategies.

However, existing task formulations and affective frameworks are not directly applicable to ToDs due to their limited control over emotion, lack of formal decision modelling, and absence of task-emotion integration. To address these gaps, we formalise the three emotional abilities specifically for ToDs and integrate them systematically into the training and interaction processes of ToD systems. Through this integration, machines can move beyond purely functional exchanges toward emotionally attuned collaboration, dynamically adapting their strategies to promote smoother, more effective human-machine interaction.

The next chapter builds upon this foundation by examining dialogue theory and the task-oriented dialogue pipeline, into which these affective capabilities can be systematically incorporated. In doing so, we connect the theoretical insights of affective computing to the concrete mechanisms of dialogue management, ultimately advancing toward the design of emotionally intelligent task-oriented dialogue systems.

## Chapter 4

# Dialogue Theory

### 4.1 Overview

In this chapter, we establish the theoretical and computational foundations of dialogue in order to situate our study of task-oriented dialogue (ToD). ToD is distinctive in its orientation towards well-defined user goals, the factual information retrieval from a database, and the collaborative processes that enable their achievement. While ToD is the central concern of this thesis, it is also important to consider other dialogue settings, especially chit-chat. Methods and resources developed for these domains, particularly for modelling emotion, can inform and enrich ToD research.

To make sense of these diverse settings, it is necessary to look beyond surface differences and consider the theoretical foundations of dialogue. Dialogue is more than the transaction of information; it is also a form of action. Speech Act Theory (Austin, 1962) frames utterances as acts (e.g. inform, request, confirm), a perspective that directly informs dialogue act annotation in ToDs, whereas in chit-chat, speech acts are often phatic or affective.

Grice's Cooperative Principle (Grice, 1975) emphasises collaboration, which in ToD is realised through user-system cooperation in achieving goals, and in chit-chat, cooperation sustains flow and rapport.

Clark's theory of grounding (Clark and Brennan, 1991) highlights the establishment of shared understanding through interaction. In ToDs, this process is supported by mechanisms such as confirmation and repair, supported by dialogue state tracking (DST) and database query. In chit-chat, minimal acknowledgements often suffice to maintain conversational alignment.

Finally, partially observable Markov decision processes (POMDPs) formalise ToD as sequential decision-making, providing a principled and computational framework for dialogue management (Young, 2002). The explicit goals, a bounded domain, and measurable outcomes such as task success in ToD offer a well-defined objective for optimisation. The imperfect information from speech recognition, intent inference, and user behaviour makes POMDPs particularly suitable.

These perspectives frame dialogues as actionable, collaborative, grounded, and optimisable. They provide the conceptual and technical basis for ToD while clarifying how it differs from chit-chat. The fundamental differences between ToDs and chit-chat motivate the research presented in this thesis, including dataset construction, the development of specialised emotion recognition techniques, and the systematic integration of emotion throughout the ToD pipeline.

### 4.2 Core Theories of Dialogue

The study of dialogue has been shaped by several foundational theories in linguistics, philosophy, and cognitive science, many of which have influenced the design of computational dialogue systems. This section outlines the most relevant of these theories: speech act theory, the cooperative principle, and the theory of common ground, before turning to computational models that formalise task-oriented dialogues as a decision-making process. Together, these theories provide a multi-layered view of how dialogues function and how they motivate the modelling of task-oriented dialogues.

### 4.2.1 Speech Act Theory

A seminal contribution to the philosophy of language is *Speech Act Theory* (Austin, 1962), which argues that utterances are not merely strings of words conveying propositional content, but actions performed in a social context. In this view, to speak is to act. For example, when a speaker says, “Can you book me a table?”, they are not merely describing a state of the world but performing the act of requesting.

Searle (1969) refined Austin’s original account by distinguishing three layers of speech acts:

- **Locutionary act:** the act of producing a meaningful utterance (e.g., the literal sentence “Can you book me a table?”).
- **Illocutionary act:** the speaker’s communicative intent (e.g., making a request).
- **Perlocutionary act:** the effect of the utterance on the hearer (e.g., persuading the hearer to take action).

This layered view has been especially influential for dialogue research, since computational systems need to infer not only what was said (locution) but what was meant (illocution), and sometimes also anticipate the expected effect (perlocution). Early computational dialogue models drew directly from speech act theory by formalising *dialogue acts*, categories that map onto illocutionary functions. Dialogue acts such as *request*, *inform*, *confirm*, and *reject* provide the functional vocabulary through which systems interpret user input and select appropriate responses. Formal theories and annotation schemes proposed by Traum (1995) and Bunt (2011) established dialogue acts as the backbone of computational dialogue modelling.

**Dialogue Acts in Different Scenarios** In ToDs, dialogue acts form the building blocks of dialogue managers, defining the communicative functions that guide the system’s behaviour towards goal completion. For instance, a restaurant booking system may rely on *request(food\_type)*, *inform(area)*, or *confirm(day)* to track and progress towards the user’s goal. In chit-chat, by contrast, dialogue acts are less about achieving a concrete outcome and more about maintaining social presence, often through phatic or affective acts such as greetings, acknowledgements, or expressions of empathy.

Thus, speech act theory provides both a theoretical lens for analysing human communication and a practical foundation for computational dialogue systems. Its operationalisation through dialogue acts continues to underpin dialogue modelling, particularly in ToD, while also offering insight into how communicative functions vary across different dialogue scenarios.

**Emotional Aspects of Dialogue Acts** Emotional meaning often intersects with the functional layer of speech acts. Emotions can shape how an act is performed (e.g., an angry request vs. a polite one), signal relational stance (e.g., empathy through acknowledgements or apologies), and influence how the hearer interprets intent or chooses a response. In this sense, emotion facilitates mutual understanding by modulating the illocutionary force and perlocutionary effect of speech acts. In computational dialogue systems, this has motivated extensions of dialogue act taxonomies to include emotional aspects (Bunt et al., 2020), where communicative functions such as comforting, praising, or apologising explicitly encode both intent and affect. Such integration of emotional and functional layers provides a richer foundation for affect-aware dialogue management, as later explored in this thesis.

### 4.2.2 The Cooperative Principle

Grice (1975) introduced the *Cooperative Principle*, which views conversation as a fundamentally collaborative activity. According to this principle, interlocutors assume that contributions are made in good faith to advance the purposes of the exchange. Grice articulated this through four conversational maxims:

- **Quality:** be truthful, do not say what you believe to be false.
- **Quantity:** be as informative as required, but not more than necessary.
- **Relation:** be relevant to the conversational goal.
- **Manner:** be clear, brief, and orderly.

Although these maxims are frequently disregarded for rhetorical or social purposes (e.g., irony, understatement, humour), they establish the baseline expectations that make dialogue intelligible. Participants interpret each other's utterances under the assumption that these principles are being respected, and they can re-interpret apparent violations as purposeful deviations.

**Conversational Cooperation in Different Scenarios** In ToDs, cooperation takes the form of joint problem-solving between the user and system. The system must provide accurate, relevant, and clear information, while the user contributes by clarifying goals, rephrasing queries after misunderstandings, and responding to system prompts. Computationally, this principle underpins strategies for error recovery, clarification requests, and the balance between system-initiative and user-initiative dialogue.

In chit-chat, cooperation does not aim at completing a well-defined task but rather at maintaining mutual engagement, rapport, and social presence. Here, maxims are interpreted more flexibly: relevance may be oriented towards sustaining flow rather than topical focus, and quantity may be adapted to maintain rhythm and affective connection.

The cooperative nature of dialogue has also been formalised in computational terms. The theory of joint intentions (Cohen and Levesque, 1990) models conversation as the alignment of agents' goals and commitments, providing a basis for collaborative dialogue management in AI systems. This formalisation connects Grice's philosophical insights to practical mechanisms for representing, tracking, and updating shared goals in dialogue systems.

**Emotion in Conversational Cooperation** Emotion plays an important role in sustaining the cooperative nature of conversation. While the Cooperative Principle focuses on rational adherence to conversational maxims, emotion provides the affective signals that help interlocutors interpret, regulate, and sometimes strategically violate those norms. For instance, emotional tone often conveys the speaker's stance toward the maxim of quality (sincerity), where anger or sarcasm may indicate deliberate flouting. Positive affect, such as enthusiasm, can enhance the relationship by reinforcing engagement and shared purpose, while frustration or confusion may alert the listener to a breakdown in manner or quantity (unclear or insufficient information). In this sense, emotion facilitates cooperation not through propositional content, but by signalling alignment, attentiveness, and willingness to maintain mutual understanding. Computationally, modelling these affective cues enables dialogue systems to monitor the emotional aspect of interaction and adapt their cooperative behaviour accordingly by, for example, clarifying when frustration is detected, softening tone when user irritation rises, or expressing empathy to restore trust and collaboration. In this thesis, this view of emotion as a facilitator of cooperative communication informs the design of emotion modelling and affect-aware dialogue management, where emotional cues are treated as signals of user engagement and task alignment within task-oriented interaction.

### 4.2.3 Grounding in Conversations

The notion of *grounding* was developed by Clark and Brennan (1991), who argued that dialogue is a continual process of establishing and maintaining *common ground*: the set of shared beliefs, assumptions, and knowledge that interlocutors rely on to ensure mutual understanding. In this framework, every conversational contribution carries a dual responsibility: the speaker must provide information, and the hearer must demonstrate understanding. Evidence of understanding may

take the form of acknowledgements (e.g., “okay”), clarifications (e.g., “do you mean tonight or tomorrow?”), or repair mechanisms when misunderstandings occur. Without these signals, dialogue would lack the feedback loop necessary to ensure successful coordination.

Grounding has been central to computational dialogue modelling. Traum (1999) formalised mechanisms of acknowledgement, repair, and confirmation, showing how they can be encoded to maintain alignment in collaborative systems. These ideas directly inform the architecture of modern dialogue systems.

**Grounding in Different Dialogue Scenarios** In ToDs, grounding is achieved through interactional actions such as explicit confirmations, acknowledgements, and repair strategies, which provide evidence that both participants accept the information as shared (Mohapatra et al., 2024). These grounding actions are supported computationally by DST, which maintains the system’s evolving belief about user goals, enabling the system to determine when confirmation or repair is required. Database queries further contribute to grounding by providing verifiable task outcomes, allowing both participants to align on entities or constraints once the system retrieves and presents consistent results.

In chit-chat, grounding tends to be lighter and more implicit. Minimal cues such as backchannels (“uh-huh”, “I see”), alignment markers (“exactly”), or laughter are often sufficient to maintain conversational flow. These cues may not carry propositional content but function to reassure the speaker that they have been understood and that interactional alignment is intact.

Taken together, grounding theory highlights dialogue as a dynamic process of alignment, where interlocutors continually negotiate shared meaning. In computational settings, this insight has motivated explicit representations such as dialogue states and database queries in ToD, while offering explanatory power for the more lightweight or affective grounding practices found in chit-chat and interviews.

**Emotion and Grounding** Emotion is deeply intertwined with the process of grounding, as affective cues serve to build and reinforce mutual understanding. While grounding theory focuses on the exchange of informational evidence, emotion provides social evidence of alignment, such as attentiveness, agreement, or concern, through tone, prosody, and facial expression. Positive affect, such as warmth or empathy, can strengthen the sense of shared understanding, while negative emotions like frustration or confusion often signal breakdowns that require repair. In this way, emotion facilitates grounding by making understanding perceptible: it allows interlocutors to assess not only what has been comprehended, but also how it has been received. Computationally, incorporating emotional feedback into grounding mechanisms enables dialogue systems to detect when the user feels misunderstood, adapt confirmation strategies, and express empathy to restore mutual alignment (Picard, 1997). This perspective motivates our work in this thesis, which models emotion as a dynamic cue that reflects the establishment or breakdown of factual grounding, particularly around database query results.

#### 4.2.4 Partially Observable Markov Decision Process

While speech act theory, cooperation, and grounding offer conceptual insights into dialogue, computational dialogue research has also sought formal frameworks for managing interaction. A particularly influential paradigm is to treat dialogue as a sequential decision-making problem under uncertainty, captured by the formalism of *Partially Observable Markov Decision Processes (POMDPs)*.

Formally, a POMDP is defined by the tuple:

$$\mathcal{M} = \langle S, \mathcal{A}, \mathcal{O}, T, Z, R \rangle \quad (4.1)$$

where:

- $\mathcal{S}$  is a set of hidden states that capture all factors relevant to the decision process.
- $\mathcal{A}$  is the set of possible actions available to the agent.
- $\mathcal{O}$  is the set of possible observations that provide indirect evidence of the underlying state.
- $T(s' | s, a)$  is the state transition model, i.e., the probability of moving to  $s'$  given current state  $s$  and action  $a$ .
- $Z(o | s')$  is the observation model, i.e., the probability of receiving observation  $o$  given that the state is  $s'$ .
- $R(s, a)$  is the reward function, encoding the utility of taking action  $a$  in state  $s$ .

Because the underlying state  $s$  is not directly observable, the agent maintains a *belief state*  $b(s)$ , a probability distribution over states. When action  $a$  is taken and observation  $o$  is received, the belief state is updated according to Bayes' rule:

$$b'(s') = \eta Z(o | s') \sum_{s \in \mathcal{S}} T(s' | s, a) b(s), \quad (4.2)$$

where  $\eta$  is a normalisation constant. Decision-making is then framed as selecting an action that maximises the expected long-term return with discount factor  $\gamma \in [0, 1)$ :

$$\pi^*(b) = \arg \max_{a \in \mathcal{A}} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid b_0 \right]. \quad (4.3)$$

**Modelling Dialogues with POMDP** The mathematical formalisation of POMDP provides a principled way to model sequential interaction under uncertainty, where the agent must reason about hidden states, update beliefs as new observations arrive, and choose actions that optimise expected outcomes over time. Therefore, POMDPs are particularly well-suited to model **task-oriented dialogue**. Young (2002) adapted the POMDP framework to dialogue management, with later surveys consolidating its applications (Young et al., 2013). Here, the state space corresponds to user goals and slot values, the observation model accounts for errors in speech recognition and natural language understanding, and the reward function encodes task success (e.g., booking completed), balanced with efficiency (shorter dialogues). By optimising policies under uncertainty, POMDP-based systems embody both the cooperative and goal-driven nature of ToD.

In **chit-chat**, however, POMDPs are less applicable. Open-ended goals, loosely defined success criteria, and the absence of a clear reward structure make it difficult to specify  $R(s, a)$  or even define a bounded state space.

**POMDP with Emotion** While classical POMDP-based dialogue management focuses on uncertainty about user goals and intents, real human dialogue involves additional affective uncertainty. User emotions influence how utterances are produced, interpreted, and evaluated, and thus form part of the latent state the system must infer. Extending the standard formulation, the affective POMDP can be defined as:

$$\mathcal{M}_e = \langle \mathcal{S} \times \mathcal{E}, \mathcal{A} \times \mathcal{A}_e, \mathcal{O}_u \times \mathcal{O}_e, T_e, Z_e, R_e \rangle. \quad (4.4)$$

where, on top of Equation 4.1:

- The original hidden state  $\mathcal{S}$  is extended with the space of user emotional states  $\mathcal{E}$ .
- The original action space  $\mathcal{A}$  is combined with affective actions  $\mathcal{A}_e$ .
- The observation space  $\mathcal{O}$  contains two aspects:  $\mathcal{O}_u$  denotes semantic observations and  $\mathcal{O}_e$  denotes affective observations, such as linguistic cues and sentiment.

- $T_e(s', e' | s, e, a, a_e)$  models the transition dynamics over both semantic and emotional states.
- $Z_e(o_u, o_e | s', e')$  captures the joint likelihood of observing semantic and emotional signals.
- $R_e(s, e, a)$  is the reward that incorporates both task success and affective outcomes, such as user satisfaction or trust.

The belief state now tracks a joint distribution over task and emotional variables under the new equation

$$b'(s', e') = \eta Z(o_u, o_e | s', e') \sum_{\{s, e\} \in \{S, \mathcal{E}\}} T_e(s', e' | s, e, a, a_e) b(s, e), \quad (4.5)$$

and the optimal policy  $\pi^*(b)$  is then learned to maximise the expected affective return:

$$\pi^*(b) = \arg \max_{\{a, a_e\} \in \{\mathcal{A}, \mathcal{A}_e\}} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_e(s_t, e_t, a_t, a_{e,t}) \mid b_0 \right]. \quad (4.6)$$

This formulation treats emotion as both an observable signal (via  $\mathcal{O}_e$ ) and a latent factor influencing user goals and cooperation. For example, rising frustration may increase the probability of misunderstanding (affecting  $T_e$ ), while positive affect may reinforce belief in successful grounding or task completion. Optimising  $R_e$  thus encourages the agent to select semantic and affective actions that maintain not only task efficiency but also affective alignment, producing more cooperative and engaging dialogue behaviour. This view motivates our work on accurate user emotional state estimation (Chapter 5, 6 and 9) and the incorporation of the full affective POMDP into the task-oriented dialogue system pipeline (Chapter 8 and 9).

## 4.3 Dialogue Settings

### 4.3.1 Task-oriented Dialogues

ToDs are interactions in which an operator, either a human or an artificial intelligence agent, help a user to accomplish a specific, well-defined goal. Classic examples include booking a restaurant, arranging travel, or troubleshooting a technical problem. What makes ToD distinctive is its structural clarity: the user typically enters the dialogue with a concrete task in mind, the interaction is confined to a constrained domain such as restaurants or hotels, and successful completion depends on retrieving factual information from a structured database or knowledge base. These characteristics allow ToD to be modelled with a precision and formality that is less accessible in open-ended or therapeutic conversations.

Each of the core dialogue theories outlined earlier finds a natural correspondence within this setting. Speech Act Theory provides the foundation for dialogue acts, which operationalise communicative intent in ToD. Utterances are not mere statements of fact but actions such as requests, confirmations, or rejections, each of which drives the dialogue forward towards the user's goal. The Cooperative Principle is equally central. ToD is inherently collaborative: the system must present information truthfully, relevantly, and clearly, while the user cooperates by clarifying their requirements, confirming retrieved results, or reformulating queries when miscommunication occurs. In this sense, both parties jointly contribute to successful task completion, embodying Grice's insight that conversation is guided by shared norms of cooperation. Grounding further underpins this collaboration. As the dialogue progresses, the user and the system must establish and maintain a common understanding of constraints and commitments. Computationally, this is realised through a series of grounding actions such as confirmation and repair. Such actions are supported by dialogue state tracking, which records the evolving set of agreed slot-value pairs, and symbolic database queries, which anchor the dialogue in an external resource.

Computationally, the decision-making aspect of ToD has been formalised through Partially Observable Markov Decision Processes (POMDPs). This framework accounts for the inherent uncertainty

of speech recognition and intent inference, modelling dialogue as sequential decision-making under partial observability, and optimising dialogue policies for long-term task success. Specifically, in a POMDP  $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, R \rangle$ ,

- $\mathcal{S}$  is the set of possible dialogue states (e.g., user goals, filled slots).
- $\mathcal{A}$  is the set of system actions (e.g., *request*, *inform*).
- $\mathcal{O}$  is the set of possible observations (e.g., user utterances and noisy interpretation).
- $T(s' | s, a)$  is the transition model defining the probability of moving to state  $s'$  from state  $s$  after action  $a$ .
- $Z(o | s)$  is the observation model capturing uncertainty in interpreting user input.
- $R(s, a)$  is the reward function, encoding task success and efficiency.

The dialogue state is not directly observable. Therefore, the system maintains a *belief state*  $b(s)$ , a probability distribution over all possible dialogue states. The implementation of task-oriented dialogue systems have been discussed in Section 1.1.

Although ToD is primarily factual and goal-oriented, emotion plays a subtle but important role in shaping interaction quality (Walker et al., 1997). Users frequently express frustration when systems fail to understand their input or when repeated clarification is required (Skantze, 2007), and such negative affect can undermine the overall experience even if the task is ultimately completed. Conversely, systems that can recognise and adapt to user emotions tend to foster trust and increase user satisfaction (Bickmore and Picard, 2005). Such an ability can be demonstrated by detecting frustration, offering empathetic responses, or adjusting strategies to reduce user effort. Emotion in ToD is therefore best understood not as an additional task to be solved, but as a moderating factor that influences the perceived success and effectiveness of goal-driven dialogue.

In summary, task-oriented dialogue exemplifies a structured conversational form that is grounded in clear goals, bounded domains, and database-driven factual retrieval. Its tight alignment with theoretical perspectives such as speech acts, cooperation, grounding, and decision-theoretic formalisation makes it both conceptually robust and computationally tractable. At the same time, its sensitivity to users' emotional states highlights the need to extend traditional models with affective intelligence, thereby combining efficiency with responsiveness to human needs.

### 4.3.2 Chit-chat Dialogues

Chit-chat dialogues, sometimes referred to as social or open-domain dialogues, differ fundamentally from task-oriented exchanges in that they are not anchored to a specific, predefined goal. Instead, their purpose is largely relational: they help build rapport, maintain social connections, and enable the expression of personality or affect. Conversations may drift across topics with little structure, and success is not measured by the completion of a task but by the quality of engagement, mutual enjoyment, and continuity of interaction. This open-endedness makes chit-chat both a common and essential part of human communication, but also a particularly challenging setting for computational modelling, since the space of possible topics and responses is virtually unbounded.

Theories of dialogue illuminate how chit-chat functions despite its lack of concrete goals. Speech Act Theory is relevant here insofar as the dominant acts are phatic and affective rather than instrumental. Utterances such as greetings, acknowledgements, or expressions of empathy do not advance a specific task but sustain the interaction and reinforce social bonds. The Cooperative Principle also applies, though cooperation in chit-chat is directed not at problem-solving but at maintaining conversational flow and positive affect. Speakers respect Grice's maxims in a loose sense, being relevant and clear enough to sustain engagement, yet may also playfully flout them through humour, exaggeration, or irony, without breaking the cooperative frame. Grounding in chit-chat is lightweight

compared to more task-driven dialogues. Minimal acknowledgements, backchannels, and alignment markers (“I know”, “exactly”, laughter) are typically sufficient to establish mutual understanding. There is little need to maintain a detailed common ground beyond the immediate conversational context, since the interaction is not oriented towards cumulative task completion.

Computation-wise, chit-chat dialogues are frequently modelled as a sequence-to-sequence generation problem, where language models or neural dialogue systems generate responses directly from conversational history without an explicit state representation or reward-driven optimisation framework. This approach reflects the fluidity of social conversation, but also highlights the difficulty of defining coherence, relevance, and engagement in the absence of formal success criteria.

Emotion, by contrast, is central to chit-chat. Indeed, affective expression and perception are often the primary means by which the dialogue achieves its purpose. Systems that can recognise and generate emotional cues, through word choice, prosody, or dialogue strategies, are more likely to sustain engaging and human-like interaction. The absence of task constraints also makes emotional sensitivity more salient: a chit-chat system that fails to respond appropriately to affective signals risks breaking the illusion of naturalness, since social presence and empathy are the core metrics by which such systems are judged.

In sum, chit-chat dialogues illustrate the socially grounded nature of human communication. They show how dialogue can function without a predefined goal, guided instead by affective expression, phatic acts, and lightweight grounding. For computational systems, chit-chat presents a distinct set of challenges: the lack of formal structure makes modelling difficult, yet the centrality of emotion makes it a crucial testing ground for affect recognition and generation. Compared to task-oriented dialogues, chit-chat highlights the social dimension of interaction, providing an essential contrast for understanding how emotion might enrich goal-driven systems.

### 4.3.3 Comparative Perspective

The two dialogue settings considered in this chapter, task-oriented dialogues and chit-chat dialogues, illustrate the diversity of conversational forms. Each type of dialogue is structured by a different set of goals, interactional norms, and evaluative criteria, which in turn shape how the core theories of dialogue manifest in practice.

Dialogue Setting	Dialogue Acts	Source of Grounding	Collaborative Nature	Computational Modelling	Role of Emotions
<b>Task-oriented Dialogue</b>	Instrumental acts (request, inform, confirm, etc.)	Explicit confirmation, repair	Strongly collaborative for task success	POMDP-based decision making	Moderating factor affecting satisfaction and efficiency
<b>Chit-chat or Open-domain Dialogue</b>	Phatic and affective acts (greetings, empathetic, etc.)	Backchannels, light confirmations	Loosely collaborative for sustaining flow and rapport	Sequence-to-sequence generation or language modelling	Central to engagement and rapport; expressed through empathy and agreement

FIGURE 4.1: A comparison of dialogue settings across theoretical, collaborative, and computational dimensions, with emphasis on the role of emotion.

Figure 4.1 summarises the key similarities and differences across dialogue settings. While both rely on fundamental dialogue theories such as speech acts, cooperation, and grounding, they operationalise them in distinct ways depending on their goals and structures. Task-oriented dialogues emphasise efficiency and factual accuracy, supported by formal state tracking and decision-making models, with emotion acting as a moderating influence. In contrast, chit-chat dialogues prioritise

relational engagement, drawing on phatic and affective acts with lightweight grounding, where emotion is central to sustaining rapport.

This comparative perspective motivates the focus of this thesis. Because task-oriented and chit-chat dialogues differ fundamentally in goals and interactional structure, emotional intelligence cannot be transferred directly from one setting to the other. Task-oriented and chit-chat dialogues differ fundamentally in their goals and interactional structure. Therefore, despite emotion having been extensively studied in chit-chat interactions, emotional intelligence cannot be directly transferred from one setting to the other. Instead, emotion must be studied and integrated specifically within the constraints and objectives of ToD. While insights from chit-chat research provide valuable intuition about affective communication, ToD systems require dedicated models, datasets, and integration strategies that account for goal-driven interaction and structured decision-making. The research presented in this thesis, therefore, investigates how emotion can be systematically incorporated into ToD systems while preserving their functional requirements.

#### 4.3.4 Other Dialogue Settings

Beyond task-oriented and chit-chat dialogues, a range of specialised conversational settings have been studied in conversational AI research, each characterised by distinct interactional goals, participant roles, and evaluation criteria.

**Psychological interviews** are commonly encountered in clinical, counselling, and assessment contexts, where the objective is to elicit, interpret, and understand personal experiences and emotional states. The interaction is typically asymmetric: the interviewer guides the conversation through structured or semi-structured questioning, while the interviewee provides reflective and emotionally grounded responses. Such dialogues aim to enable reliable understanding and assessment of an individual's mental state, including diagnostic inference and affective analysis (Gratch et al., 2014). Psychological interviews therefore, provide a controlled setting for studying nuanced expressions of human affect and mental state. In Chapter 7, we evaluate the ability of large language models to identify mental distress from psychological interviews, in addition to task-oriented and chit-chat dialogues.

**Pedagogical interactions** represent another important dialogue setting, in which conversational agents act as tutors or learning companions. Tutoring dialogues are designed to facilitate learning through explanation, guided questioning, feedback, and formative assessment (Dinucu-Jianu et al., 2025; Saeki et al., 2024). Unlike task-oriented systems that optimise task completion, pedagogical dialogue systems are evaluated based on learning outcomes, student engagement, knowledge retention, and assessment accuracy. These interactions often involve adaptive strategies that respond to a learner's knowledge state and misconceptions instead of providing the solution straightaway.

Dialogue agents have also been widely explored for other **social and interpersonal purposes**. Examples include attentive listening agents, conversational partners for job interviews or speed-dating practice, laboratory guides, and systems designed for emotional support conversations (Glas et al., 2016; Liu et al., 2021). In these settings, the primary objective is not task completion but sustaining socially appropriate interaction, fostering rapport, and providing emotional or motivational support. Such dialogues often require sensitivity to social cues, empathy, and long-term conversational coherence, making them particularly relevant for studying socially aware conversational behaviour.

These dialogue settings, while not exhaustive, highlight that conversational interactions can differ substantially along several dimensions, including interaction symmetry, communicative goals, and evaluation criteria. While ToDs prioritise efficient goal completion and chit-chat emphasises engagement, psychological interviews focus on affective understanding, pedagogical dialogues target knowledge acquisition and assessment, and socially oriented interactions prioritise rapport and interpersonal support. These differences imply distinct requirements for dialogue modelling, such as reasoning about user mental states, adapting to learner knowledge, or maintaining socially appropriate behaviour over extended interactions. Importantly, emotion also plays different functional roles across these settings: it may serve as conversational context in social dialogue, a diagnostic signal in

psychological interviews, or a facilitator of engagement and learning in pedagogical interactions. As a result, building effective conversational agents requires an understanding of the specific interactional characteristics and emotional dynamics associated with each dialogue setting.

## 4.4 Conclusion

This chapter has established the theoretical and computational foundations of dialogue in order to situate ToDs within the broader landscape of human communication. We began by outlining three core theoretical perspectives: Speech Act Theory, which frames utterances as communicative actions; the Cooperative Principle, which highlights dialogue as a collaborative endeavour; and Grounding, which emphasises the continual establishment of shared understanding. Together with the formalism of Partially Observable Markov Decision Processes (POMDPs), these perspectives provide complementary insights into dialogue as action, collaboration, mutual alignment, and decision-making under uncertainty.

Building on these foundations, we then examined two important dialogue settings in detail. Task-oriented dialogues were shown to be goal-driven, domain-constrained, and database-centred, making them highly amenable to computational modelling and systematic evaluation. Chit-chat dialogues, by contrast, exemplify open-ended social interaction, where phatic acts, lightweight grounding, and affective engagement dominate. A comparative perspective highlighted both the commonalities across settings, such as reliance on dialogue acts, cooperation, and grounding, and the distinctive ways in which these principles are realised depending on the goals and structure of the interaction. Beyond these two paradigmatic settings, we also considered specialised dialogue contexts such as psychological interviews, pedagogical interactions, and socially oriented conversations, which differ in interactional structure, evaluation criteria, and the functional role of emotion. These settings illustrate that conversational behaviour cannot be understood through a single dialogue paradigm, as effective dialogue modelling requires sensitivity to the distinct communicative goals and emotional dynamics inherent to each interaction type.

The chapter as a whole underscores two key insights. First, ToD stands out as the most tractable setting for computational modelling, benefiting from clear goals, structured domains, and well-defined success criteria. Second, insights from other dialogue types reveal the limits of a purely goal-oriented view: chit-chat highlights the centrality of emotion, while other settings highlight the necessity of rapport, empathy, and the interpretation of specific affective states. These comparisons motivate the central argument of this thesis: that emotionally intelligent ToD systems must combine the formal rigour of computational models with sensitivity to the affective and relational dynamics that characterise human dialogue.

## Chapter 5

# EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems

This chapter summarises our work on constructing a large-scale, emotion-annotated corpus for task-oriented dialogues enabling fine-grained user emotion recognition and gives a verbatim copy of our paper (Feng et al., 2022):

Shutong Feng et al. (June 2022). “EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4096–4113. URL: <https://aclanthology.org/2022.lrec-1.436/>

## 5.1 Summary

Despite the crucial role of emotion in human communication, user emotions have long been neglected in task-oriented dialogue systems, where the primary focus has traditionally been on task success. However, user emotion plays a vital role in task-oriented interactions: it serves as implicit feedback on the system’s performance and contributes to more natural, human-like interactions.

A major barrier to advancing research in this area is the lack of appropriate resources. Existing datasets are either limited in scale or lack annotations that capture the nuanced emotional states typical of task-oriented scenarios. To address these issues, we construct EmoWOZ, a large-scale corpus with manually annotated user emotions, specifically designed for task-oriented dialogues.

EmoWOZ includes a diverse mix of both human-human and human-machine conversations. Its annotation scheme is grounded in the Ortony Clore and Collins (OCC) cognitive theory (Ortony et al., 1988) of emotion but adapted to the unique characteristics of task-oriented dialogue. Using this dataset, we benchmark a variety of generic emotion recognition models, demonstrating the inherent difficulty of detecting user emotions in goal-directed interactions.

Furthermore, we show that integrating emotion recognition into a dialogue state tracker can lead to improved performance. This highlights the broader potential of emotion modelling, not just as an end in itself, but as a means to enhance other components of task-oriented dialogue systems.

## 5.2 Personal Contributions

The implementation, technical results, and writing are my own work. Michael Heck contributed to the experiment to integrate emotion recognition into dialogue state tracking. The human-machine dialogues in EmoWOZ (DialMAGE) are collected using the dialogue system built by Christian Geishauer, Hsien-chin Lin, and Carel van Niekerk. Milica Gasić supervised the project. Nunu Lubis co-supervised the project. Co-authors assisted in writing and proofreading.

# EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems

Shutong Feng, Nurul Lubis, Christian Geishauer, Hsien-chin Lin,  
Michael Heck, Carel van Niekerk and Milica Gašić

Heinrich Heine University Düsseldorf  
Universitätsstraße 1, 40225 Düsseldorf, Germany  
{fengs, lubis, geishaus, linh, heckmi, niekerk, gasic}@hhu.de

## Abstract

The ability to recognise emotions lends a conversational artificial intelligence a human touch. While emotions in chit-chat dialogues have received substantial attention, emotions in task-oriented dialogues remain largely unaddressed. This is despite emotions and dialogue success having equally important roles in a natural system. Existing emotion-annotated task-oriented corpora are limited in size, label richness, and public availability, creating a bottleneck for downstream tasks. To lay a foundation for studies on emotions in task-oriented dialogues, we introduce EmoWOZ, a large-scale manually emotion-annotated corpus of task-oriented dialogues. EmoWOZ is based on MultiWOZ, a multi-domain task-oriented dialogue dataset. It contains more than 11K dialogues with more than 83K emotion annotations of user utterances. In addition to Wizard-of-Oz dialogues from MultiWOZ, we collect human-machine dialogues within the same set of domains to sufficiently cover the space of various emotions that can happen during the lifetime of a data-driven dialogue system. To the best of our knowledge, this is the first large-scale open-source corpus of its kind. We propose a novel emotion labelling scheme, which is tailored to task-oriented dialogues. We report a set of experimental results to show the usability of this corpus for emotion recognition and state tracking in task-oriented dialogues.

**Keywords:** Emotion Recognition in Conversations, Task-oriented Dialogues

## 1. Introduction

Incorporating human intelligence into conversational artificial intelligence (AI) has been a challenging and long-term goal (Picard, 1997). Emotional intelligence, defined as the ability to regulate, perceive, assimilate, and express emotions, is a key component of general intelligence (Mayer et al., 1999). Such emotion awareness can help the conversational AI generate more emotionally and semantically appropriate responses (Zhou et al., 2018).

Dialogue systems generally fall into two classes. Task-oriented systems converse with users to help complete tasks determined by user goals. Chit-chat systems are set up to mimic the unstructured conversations or ‘chats’ characteristic of human-human interaction (Jurafsky and Martin, 2009). Chat-oriented systems are typically modelled in a supervised fashion with large available corpora (Vinyals and Le, 2015). In contrast, task-oriented systems track the user goal throughout the dialogue and a policy is typically trained via some form of reinforcement learning (RL) to conduct dialogue towards successful goal completion (Young, 2002). Moreover, the scope of the dialogue can also be extended during this process, e.g. by adding new domains to the dialogue system (Madotto et al., 2021). Consequently, the distribution of data from which a task-oriented system learns can change.

Emotions appear in both chit-chat and task-oriented dialogues. However, the cause of emotion may differ as well as their role. Chit-chat dialogues are a means to express emotion. Speakers may discuss emotional ex-

periences (Li et al., 2017), or topics that induce emotions such as news broadcasts (Lubis et al., 2017). In task-oriented dialogues, the user is primarily interested in achieving their goal. While an emotional situation may be a reason to interact with the system, e.g. the user just missed a flight and needs to rebook one, the emotion the user exhibits is more often a reaction to potential goal completion or failure. Since the emotion is centred around the user goal, it is more contextual and subtle. Therefore, besides inferring emotional states from dialogue utterances, an agent also needs to reason about emotion-generating situations (Poria et al., 2021).

Substantial research efforts in emotion recognition in conversations (ERC) have been invested in chit-chat dialogues (Majumder et al., 2019; Ghosal et al., 2020). There are several public ERC corpora containing chit-chat dialogues (Li et al., 2017; Poria et al., 2019; Zehri and Choi, 2018) and conversational data from social media (Zhou and Wang, 2018). These corpora can tremendously accelerate the building of emotional chatbots using data-driven approaches (Zhou et al., 2018). In task-oriented dialogues, recognising emotions is equally important but remains largely unaddressed. Using RL to optimise a dialogue policy necessitates a feedback signal. While it is accepted that the feedback signal needs to correlate with user satisfaction (Ultes et al., 2017), this feedback signal is often based on hand-coded rules. Could an emotional model instead be directly used to provide such a feedback signal? Could it also be used to support emotion-aware

natural language generation (Mairesse and Walker, 2007), or even improve dialogue state tracking through multi-task learning (Heck et al., 2020a)? Existing corpora are small in size, and labels are limited to sentiment polarity, creating a bottleneck, so these questions remain largely unexplored.

In this work, we present **EmoWOZ**, a large-scale manually labelled corpus for emotion in task-oriented dialogues. EmoWOZ is derived from MultiWOZ (Budzianowski et al., 2018), one of the largest multi-domain corpora and the benchmark dataset for various dialogue modelling tasks, from dialogue state tracking (Heck et al., 2020b; Lin et al., 2021) to policy optimisation (He et al., 2022). We also collected and annotated human-machine dialogues as a complement. Our contributions are as follows:

- We construct a corpus containing task-oriented dialogues with emotion labels, comprising more than 11K dialogues and 83K annotated user utterances. To the best of our knowledge, this is the first large-scale open-source corpus & code<sup>1</sup> for emotion recognition in task-oriented dialogues.
- We propose a novel labelling scheme, containing 7 emotion classes, adapted from the Ortony, Clore and Collins (OCC) model (Ortony et al., 1988), specifically tailored to capture an array of emotions in relation to user goals in task-oriented dialogue.
- We report a series of emotion recognition baseline results to show the usability of this corpus. We also empirically show that the emotion labels can be used to improve the performance of other task-oriented dialogue system modules, in this case, a dialogue state tracker (DST).

## 2. Related Work

### 2.1. Emotion Models

Within the area of affective computing, emotion models are commonly grouped into two types: dimensional models and categorical models.

**Dimensional models** describe emotions as a combination of values across a set of dimensions. The longest established dimensions are valence and arousal, as proposed by Russell (1980) in the circumplex model of emotion. Valence measures the positivity, while arousal measures the activation. Happiness, for example, is an emotion with positive valence and high activation. Additional dimensions, namely dominance and expectancy (Fontaine et al., 2007), have also been proposed to further describe and distinguish complex emotions.

**Categorical models** group emotions into distinct categories. The “Big six” theory is one of the most well-known theories on universal emotions. Based on studies of facial expressions, Ekman (1992) proposed six

basic human emotions which are influenced neither by culture nor other social influences: happiness, anger, sadness, disgust, fear, surprise. Parrott (2001) conceptualised over a hundred emotions into a tree-structured list and identified six primary emotions from it.

Ortony et al. (1988) proposed the Ortony, Clore and Collins (OCC) emotion model, which is explicitly developed for implementation in computers. In the OCC model, 22 emotion types are described as a valenced reaction to one of three cognitive elicitors: consequences of events, actions of agents, or aspects of objects. For example, *dissatisfied* is specified as disapproving of someone else’s blameworthy action. These cognitive aspects are in line with the cognitive process of a computational agent, making the OCC model suitable for building emotional artificial agents. However, the use of this model for dialogue agents is not yet widespread. In a similar spirit, Gross and Thompson (2007) formulated the process of emotion regulation as the attention, appraisal, and response originated from various situations.

Although there are corpora with real-valued annotation of multiple emotion dimensions (Preoțiuc-Pietro et al., 2016; Buechel and Hahn, 2017), researchers often focus on the valence dimension and annotate with discrete classes (Socher et al., 2013), often called sentiment polarity. Emotion datasets also consider emotions from various categorical models in the annotation scheme (Li et al., 2017; Poria et al., 2019), but some datasets have domain-specific labels. For instance, Zhou and Wang (2018) leverage common emojis in social media posts. The Topical-Chat dataset (Gopalakrishnan et al., 2019) introduces *curious to dive deeper* in addition to other basic emotions.

In this work, we propose a novel set of 7 emotions and motivate it using OCC model as the basis. We aim for this scheme to capture the cognitive context of emotions while retaining the simplicity of labels that facilitates large-scale crowd-sourcing of emotion annotations.

### 2.2. Emotion Dialogue Datasets

Early works on ERC focus on speech signals (Cowie et al., 2001; Riccardi and Hakkani-Tür, 2005; Carrión and López-Cózar, 2008). More recently, there are increasing number of text-based ERC datasets focusing on chit-chat dialogue. Chit-chat dialogue lends itself well to affective computing research due to its open-domain set-up, where conversation topics are diverse and not restricted to a particular task. One of the largest such corpora is DailyDialog (Li et al., 2017), which contains conversations between English learners on various topics ranging from relationships to money. Other similar datasets include EmoryNLP (Zahiri and Choi, 2018) and MELD (Poria et al., 2019). They contain multi-party dialogues from the TV show *Friends*. TV recordings in talk show format have also been utilised to collect emotion-rich and topic-specific dialogues (Lubis

<sup>1</sup><https://doi.org/10.5281/zenodo.5865437>

Metric	DailyDialog	MELD	EmoryNLP	DSTC1	SentiVA	TML	EmoWOZ(Ours)
Dialogue type		Chit-chat			Task-oriented		
# Dialogues	<b>13,118</b>	1,433	897	50	1,282	3,496	11,434
Total # turns	102,979	13,708	12,606	517	35,267	68,216	<b>167,234</b>
# Unique tokens	26,364	8052	8441	199	-	-	<b>28,417</b>
Avg. turns / dialogue	7.9	9.6	14.1	10.3	<b>27.5</b>	19.5	14.63
Avg. tokens / turn	<b>14.6</b>	10.4	14.3	2.3	-	-	12.78
Label type	Emo	<b>Sent, Emo</b>	<b>Sent, Emo</b>	Sent	Sent	Sent	<b>Sent, Emo</b>
# Classes	7	<b>3 and 7</b>	<b>3 and 7</b>	3	3	5	<b>3 and 7</b>
# Annotations	<b>102,879</b>	13,708	12,606	517	35,267	68,216	83,617
Neutral Samples (%)	83.1%	47.0%	<b>30.0%</b>	-	88.6%	45.7%	70.1%
# Annotators / turn	3	3	<b>4</b>	-	3	2	3
Expert Annotator?	Yes	No	No	-	No	No	No
Agreement	0.789	0.43	0.14	-	<b>0.8</b>	0.79	0.602
Open-sourced?	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	No	No	<b>Yes</b>

Table 1: Comparison of our corpus to similar corpora. Values in bold indicate the best value for each metric. For label type, “Emo” stands for emotion categories and “Sent” stands for sentiment polarities. For corpora providing both emotion and sentiment labels, agreement metrics are measured for emotion labels. DSTC1, SentiVA, and TML refer to works by Shi and Yu (2018), Saha et al. (2020), and Wang et al. (2020), respectively.

et al., 2015). Unfortunately, existing data suitable for task-oriented corpora, such as customer service chat logs, are typically not within the public domain.

There also exist a few corpora concerning the affective aspect of task-oriented dialogues. Wang et al. (2020) proposed a large-scale sentiment classification corpus containing customer service dialogues in Chinese. However, this dataset is not publicly available. Saha et al. (2020) annotated dialogues from bAbI (Bordes et al., 2017) with sentiment for policy optimisation. Since dialogues are machine-generated, it is unclear how well these emotions match real human emotions and whether sentiment on its own sufficiently captures emotional nuances in task-oriented dialogue. In a similar spirit, Shi and Yu (2018) annotated the DSTC1 dataset with user sentiment. Unfortunately, containing only 50 dialogues, the dataset is very limited in terms of coverage and application in machine learning. To summarise, existing corpora are either limited in size or not publicly available, limiting further works on emotions in task-oriented dialogue systems. Furthermore, their annotation schemes focus on sentiment polarities, overlooking the effect of goals on users’ emotional states.

### 3. Dataset Construction

#### 3.1. Task-oriented Dialogues

**MultiWOZ:** Our dataset covers the entirety of MultiWOZ, which was constructed using the Wizard-of-Oz framework (Kelley, 1984). It contains over 10k dialogues. Each dialogue was completed by two workers, each acting as the user or the operator, to achieve specified goals such as information retrieval or making reservations. There are 7 domains in total. A single dialogue or even a single turn can span multiple domains.

**Complementary Dialogues:** We envisage emotions as learning signal for dialogue system optimisation. Since emotions in task-oriented dialogue systems can be a direct effect of the user perception of the ability of the

system to fulfill their goal, the policy performance can largely influence emotion distribution. During the life span of a data-driven task-oriented dialogue system, the distributions of dialogues and emotions may change as the policy learns and improves over time. An immediate impact of such a distributional shift is the increase in the number of negative emotions due to failed dialogues during the early stages of learning. Therefore, in addition to the human wizard policy in MultiWOZ, it is important that EmoWOZ covers a variety of dialogues which represent the emotions throughout such a dialogue system life span.

We complement MultiWOZ with human-machine dialogues from a machine-generated policy (**DialMAGE**). To elicit more genuine reactions, we let subjects directly interact with a machine-generated policy instead of human wizards trying to make machine-like mistakes. We launched a dialogue interactive task on Amazon Mechanical Turk, where workers are asked to retrieve information by interacting with the learning policy. We start with a policy trained in a supervised fashion on MultiWOZ that achieved a task success rate of 55% when evaluated with the ConvLab-2 (Zhu et al., 2020) rule-based user simulator. Throughout the task, the policy learned and improved as user feedback on task success is used for further training using RL. The policy reached a final human-rated success rate of 73%. Similar to Li et al. (2020), the policy uses a recurrent neural network (RNN) based model to produce multiple actions in a single turn, followed by the ConvLab-2 template-based NLG module for response generation.

#### 3.2. Emotion Annotation Scheme

EmoWOZ focuses on user emotions rather than system ones. We believe recognising user emotions is the starting point for building emotion-aware task-oriented dialogue systems. We use the OCC model to arrive at

Elicitor	Valence	Conduct	OCC Emotion	Our Emotion	Implication of User
Operator	Positive	Polite Impolite	Admiration, gratitude, love	Satisfied, liking, appreciative Not applicable to the dataset	Satisfied with the operator because the goal is fulfilled.
	Negative	Polite Impolite	Reproach, anger, hate	Dissatisfied, disliking Abusive	Dissatisfied with the operator’s suggestion or mistake. Insulting the operator when the goal is not fulfilled.
User	Positive	Polite Impolite	Pride, gratification	Not applicable to the dataset	
	Negative	Polite Impolite	Shame, remorse, hate	Apologetic Not modelled	Apologising for causing confusion to the operator. Insulting the operator for no reason.
Events, facts	Positive	Polite Impolite	Happy-for, gloating, love, satisfaction, relief, joy	Excited, happy, anticipating Not applicable to the dataset	Looking forward to a good event (e.g. birthday party).
	Negative	Polite Impolite	Distress, resentment, hate, fears-confirmed, pity, disappointment	Fearful, sad, disappointed Not applicable to the dataset	Encountered a bad event (e.g. robbery).
NA	Neutral	Polite Impolite	NA	Neutral Not modelled	Describing situations and needs. No emotion but rude (e.g. using imperative sentences).

Table 2: Comparison between the OCC model and our labelling scheme. Emotions that do not occur in our dataset are marked as “not applicable to our dataset”. {User, negative, impolite} has too few instances and {neutral, impolite} is not strong enough to be considered as *abusive* and therefore are not modelled for now. For simplicity, emotion words in blue are used to represent each emotion category. The OCC model is illustrated in Appendix A.

specific emotion categories. For that, we consider the following aspects:

**1. Elicitor or cause:** The OCC model defines three main elicitors of emotion: events, agents, and objects. In task-oriented dialogues, events describe the situation which brings the user to interact with the system. For example, a user may be looking for a hotel for an upcoming trip or asking for the police information after a robbery. Agents are participants of the dialogue: the user and the system. Objects are equal to entities being talked about in the dialogue, such as the recommended hotel or the nearest police station. In our dataset, an object is always associated with either the operator, who proposes it, or an event, which drives the need for it. For this reason, we do not consider the object as an elicitor alone. On the other hand, within the agent category, it is important to distinguish between the user and the system. Therefore, we arrive at three elicitors for our annotation scheme: 1) the system, 2) the user, and 3) events (or facts).

**2. Valence:** In essence, the OCC model describes emotion as a valenced reaction towards an elicitor. Valence is a dimension which expresses the positivity or negativity of emotion. For example, successfully achieving a goal is likely to bring positive valence, while a misunderstanding with an agent is likely to cause negative valence. As EmoWOZ will demonstrate in a later section, valence is highly related to task success or failure, making it an important signal for a task-oriented system. We distinguish neutral and emotional utterances, and further separate emotional utterances into those with negative and positive valence.

**3. Conduct:** Conduct is not a part of the OCC model, but given the rising concern of how humans behave when interacting with virtual assistants (Cercas Curry and Rieser, 2018), we decided to include it. Conduct describes the politeness of users and is usually associated with emotional acts. Politeness can indicate the degree of valence. For example, the user can express very strong dissatisfaction through rudeness. It also helps distinguish emotions such as those associ-

ated with apology or abuse, which are both intrinsically negative.

Considering all combinations of these three aspects for annotation leads to a large number of classes. When choosing the final set of classes we were guided by whether or not a particular emotion category occurs in the database and the potential impact of that emotion category on the dialogue policy. We also carried out several trials and considered the ease of communicating to the annotator how to label such instances. We finally arrive at a set of 6 non-neutral emotion categories:

An emotion elicited by the operator is defined as *satisfied* if it is positive, and *dissatisfied* if it is negative. Positive emotion caused by an event gives us *excited*, and negative *fearful*. In terms of negative emotions expressed towards the system, we consider user conduct to distinguish between *dissatisfied* and *abusive*, since they require very different responses from the system (Curry and Rieser, 2019). In terms of the negative emotions that users may direct toward themselves, we single out *apologetic* behaviours since it features in human-human information-seeking dialogues. Emotion categories and their attributes in the above-mentioned aspects and their relation to the original OCC model are shown in Table 2.

### 3.3. Emotion Annotation Setup

We crowd-source the emotion annotation on Amazon Mechanical Turk in a controlled manner. As suggested by Carrión and López-Cózar (2008) to improve the annotation quality, workers are shown the dialogue history up to the utterance they are required to label. Each emotion category is followed by a list of emotion words that best fit into the category and an explanation. Due to the high subjectivity in the emotion annotation (Devillers et al., 2005), each dialogue is annotated by three different workers. We also implement several measures to ensure the quality of the emotion labels:

**Qualification tests:** The test contains fifteen questions, seven are straight-forward and eight are more complex. The test also serves as a tutorial. For diffi-

cult questions, hints are provided to guide the workers to identify implicit emotions and use contextual information (see Appendix B).

**Hidden tests:** We pre-label more than 1000 utterances containing obvious emotions and use them as sanity checks. The hidden tests serve as an indicator of worker reliability. If a worker scores above 80% on the hidden tests, we assume that the worker is reliable. Otherwise, the workers’ submission is subject to manual review.

**Review for outliers:** We use a simple lexicon-based recogniser and manually annotate a small batch to have an estimate of the overall emotion distribution. If the label distribution in a worker’s submissions deviates substantially from our prior belief, we mark them for manual review.

**Annotation limit:** We limit each worker to annotate at most 500 dialogues to ensure a diversity of workers and to avoid that workers adapt to our approval policy. Overall, we had 215 workers, each annotating 160 dialogues on average.

## 4. EmoWOZ Characteristics

### 4.1. Linguistic Style

Dialogues from MultiWOZ and DialMAGE differ linguistically. As seen in Table 3, DialMAGE has longer dialogues than MultiWOZ as it takes longer for the machine-generated policy to accomplish user goals. Meanwhile, users use simpler and shorter sentences when talking to a machine. Especially when the system under-performs, users are discouraged to converse with it (see sample dialogues with annotations in Appendix C). We will analyse the impact of these differences on emotion recognition in Section 5.1.3.

	MultiWOZ	DialMAGE	EmoWOZ
# Dialogues	10,438	996	11,438
# Unique tokens	27,833	3,133	28,417
Avg. turns / dialogue	13.7	24.3	14.6
Avg. tokens / user turn	11.6	5.7	10.6
Avg. unique user tokens / dialogue	57.8	36.5	55.6

Table 3: Comparison of linguistic features in EmoWOZ and its subsets.

### 4.2. Emotion Distribution

According to Table 4, the most common non-neutral emotion in EmoWOZ is *satisfied*, followed by *dissatisfied*. This is expected in task-oriented dialogues as users mainly express emotion in relation to their goals. While MultiWOZ contains more neutral utterances, it has a more diverse emotion distribution than DialMAGE. MultiWOZ contributes most *satisfied* utterances whereas DialMAGE contributes most *dissatisfied* utterances. This is in line with their respective dialogue-generating setup.

Sometimes users also express emotion to engage or provoke the operator. MultiWOZ contains more *apologetic* and less *abusive* utterances than DialMAGE, suggesting that users tend to be more polite when talking to human operators. Dialogues from MultiWOZ also contain more event-elicited emotions (*fearful* and *excited*) than DialMAGE. Users are more talkative when conversing with human operators. Users may describe a miserable situation they were experiencing, hoping to be helped and comforted. A human operator would naturally show empathy. In MultiWOZ, the operator sometimes asks if the user is alright when the user is looking for help from a robbery. When talking to machines, users tend not to express such chit-chat-style emotions due to the expected incapability of the machine to reciprocate. This indicates that an emotionally intelligent agent will allow dialogues that are emotionally richer and more nuanced, even in a task-oriented setting.

Emotion	EmoWOZ		MultiWOZ		DialMAGE	
	Count	Prop.	Count	Prop.	Count	Prop.
Neutral	58,656	70.1%	51,426	71.9%	7,230	59.8%
Fearful	396	0.5%	381	0.5%	15	0.1%
Dissatisfied	5,117	6.1%	914	1.3%	4,203	34.8%
Apologetic	840	1.0%	838	1.2%	2	0.02%
Abusive	105	0.2%	44	0.1%	61	0.5%
Excited	971	1.2%	860	1.2%	111	0.9%
Satisfied	17,532	21.0%	17,061	23.8%	471	3.9%

Table 4: Count and prop(ortion) of emotion labels.

### 4.3. Inter-annotator Agreement

We measure the inter-annotator agreement by computing Fleiss’ Kappa (Fleiss, 1971). Fleiss’ Kappa for EmoWOZ is 0.602, suggesting a substantial agreement. Fleiss’ Kappa for MultiWOZ is 0.611, higher than 0.465 for DialMAGE. Emotions in DialMAGE are more challenging to annotate because users express emotion less explicitly when they know that they are talking to a machine that does not react to emotions. Annotators often have to infer the user’s implicit emotions from dialogue history, for example, based on repetitions or misunderstanding.

Among all utterances, 72.1% see a full agreement among three annotators, 26.4% see a partial agreement, and 1.5% see no agreement. The count of each case in each subset can be found in Appendix D. Utterances for which no agreement is reached are resolved manually. Figure 1 illustrates the confusion matrix between annotators’ labels and the golden labels. Most disagreements occur between non-neutral emotions and neutral, as well as *abusive* and *dissatisfied*. A reasonable explanation is that workers adopt different valence or impoliteness thresholds when they make decisions. Note that dissatisfied is rarely confused with abusive, but rather with neutral, suggesting that the ambiguity lies in when an expression of dissatisfaction is considered

Model	Feature	Ctx.	F1 of Each Emotion in EmoWOZ							EmoWOZ		MultiWOZ		DialMAGE	
			Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mac.	Wgt.	Mac.	Wgt.	Mac.	Wgt.
BERT	BERT	No	89.8	36.2	35.1	70.4	27.5	42.9	88.8	50.1	73.5	<b>48.4</b>	<b>83.2</b>	42.7	43.8
ContextBERT	BERT	Yes	<b>92.1*</b>	30.1	<b>61.7*</b>	62.4	<b>41.7</b>	40.8	<b>89.1</b>	54.3	<b>79.7*</b>	45.1	83.1	50.0	<b>73.5*</b>
DialogueRNN	GloVe	Yes	83.5	12.7	51.4	57.7	0.0	32.7	86.4	40.1	74.6	34.1	79.2	43.2	61.2
DialogueRNN	BERT	Yes	86.9	41.3	47.5	<b>71.5</b>	25.6	39.4	87.6	52.1	75.5	44.5	81.9	51.4	60.6
COSMIC	BERT+COMET	Yes	89.8	<b>52.0*</b>	50.7	70.9	31.6	<b>44.4</b>	88.4	<b>56.3</b>	77.1	46.7	82.7	<b>57.2</b>	61.7

Table 5: Comparison of baseline models. We report the F1 for each emotion label (**Neutral**, **Fearful**, **Dissatisfied**, **Apologetic**, **Abusive**, **Excited**, **Satisfied**) on EmoWOZ as well as **Macro** and **Weighted** F1 (excluding neutral) on EmoWOZ and its subsets. “Ctx.” stands for “context”. \* indicates statistically significant difference with  $p < 0.05$  between the best and the second best values in each column. Please refer to Appendix F.1 for more detailed results.

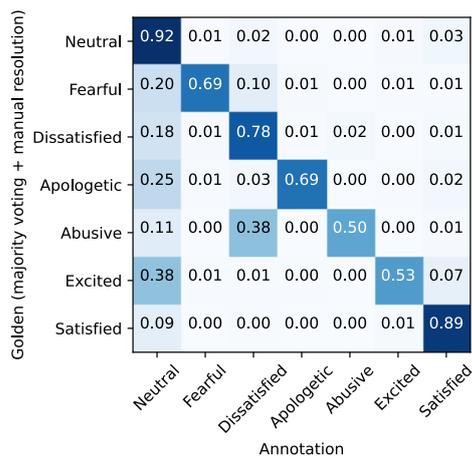


Figure 1: Confusion matrix of emotion annotations.

to be rude or abusive, and not due to the similarity between abuse and dissatisfaction.

On the other hand, confusions between *fearful* and *dissatisfied* suggest workers may also interpret elicitors differently. For example, a user may express negative emotions after the agent informed that there is no attraction meeting the user’s criteria. While the emotion is caused by the fact that there is no match, one can also argue that the operator failed to suggest alternative options. We believe differences on interpretations are natural to a certain extent, as emotion appraisal may differ across individuals (Kuppens et al., 2007).

## 5. Experiment

### 5.1. Emotion Recognition in Dialogue

Emotion recognition aims to recognise emotion within an utterance. Unlike utterances in isolation, emotion recognition in dialogues is highly contextual with respect to the dialogue history. As baselines, we compare two models originally developed for chit-chat emotion recognition as well as various BERT-based models. We believe emotion recognition is the first step towards an emotion-aware task-oriented agent, as a means for a deployed agent to obtain emotion information during an interaction.

#### 5.1.1. Baselines

**BERT (Devlin et al., 2019):** BERT is used as the utterance encoder. Each user turn is encoded in isolation without any dialogue context. The [CLS] token from a bert-base-cased model is used as the feature representation, which is then fed into a linear output layer for classification.

**ContextBERT:** The set-up is identical to that of BERT, except that the entire dialogue history and the current user utterance are concatenated in the reversed order to form one long sequence. We add “User:” and “System:” to mark the speaker of each turn.

**DialogueRNN (Majumder et al., 2019):** The model combines gated recurrent units (GRUs) with an attention mechanism to capture the long-term trajectory of the dialogue. We experiment with using GloVe embeddings (Pennington et al., 2014) or the [CLS] representation from BERT as input features. When GloVe is used, a convolutional neural network (CNN) layer is used as a feature extractor to generate utterance representations. This CNN layer is dropped when using BERT features.

**COSMIC (Ghosal et al., 2020):** This model also combines GRUs with the attention mechanism. In addition to utterance representations from a pre-trained language model (LM), it supplements input features with common-sense knowledge extracted from a pre-trained commonsense transformer model called COMET (Bosselut et al., 2019). Although the original paper uses RoBERTa as input features, we found that BERT results in a better sequence representation for emotion recognition on our data. Therefore we use BERT as the utterance encoder in our experiments.

#### 5.1.2. Experimental Setup

We perform a recognition task on the 7 emotions proposed in our annotation scheme<sup>2</sup>. All models are implemented in PyTorch (Paszke et al., 2019). For COSMIC and DialogueRNN, we use the code provided by the respective papers. We include more details on the hyperparameters of each model in Appendix E. To split EmoWOZ into training, validation, and testing sets, we

<sup>2</sup>We also performed the same experiments on 3 sentiment labels. Results can be found in Appendix F.3.

Example 1: <i>Dissatisfied</i>					Example 2: <i>Dissatisfied</i>				
<b>U:</b> I need to arrive by 15:15 <b>S:</b> I have train TR4068 leaving at 5:35 and arriving at 5:52. <b>U:</b> I want to confirm that I will arrive by 15:15? You stated, leaving at 5:35 and arriving at 5:52? [to classify]					<b>U:</b> I also need a taxi to go between the hotel and the restaurant. I'd like to leave the Gonville hotel by 09:15 <b>S:</b> When would you like to arrive by? <b>U:</b> I just mentioned that I would like to leave by 9:15 please. [to classify]				
BERT	ContextBERT	DialogueRNN (GloVe)	DialogueRNN (BERT)	COSMIC	BERT	ContextBERT	DialogueRNN (GloVe)	DialogueRNN (BERT)	COSMIC
✗ (neutral)	✓ (dissatisfied)	✓ (dissatisfied)	✓ (dissatisfied)	✓ (dissatisfied)	✗ (neutral)	✗ (neutral)	✗ (neutral)	✗ (neutral)	✗ (neutral)

Figure 2: Example dialogues and the emotion prediction for the last utterance by each model.

Training Data	Test on MultiWOZ									Test on DialMAGE								
	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mac.	Wgt.	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mac.	Wgt.
MultiWOZ	<b>95.1*</b>	<b>35.7</b>	<b>36.4*</b>	<b>70.3*</b>	<b>19.4</b>	34.1	<b>90.0</b>	<b>47.7</b>	<b>83.9</b>	80.2	<b>11.7</b>	7.7	<b>43.7</b>	11.9	<b>60.1</b>	66.3	33.6	14.5
DialMAGE	89.4	0	11.2	0	0	13.9	77.3	17.0	67.8	72.1	0	<b>75.7</b>	0	5.0	58.6	<b>71.7</b>	35.2	72.9
EmoWOZ	93.5	33.7	30.4	62.4	17.3	<b>37.1</b>	89.8	45.1	83.1	<b>81.6</b>	5.0	75.5	40.0	<b>52.8*</b>	57.3	69.2	<b>50.0*</b>	<b>73.5</b>

Table 6: Performance of ContextBERT in cross-dataset experiments. We report the F1 for each emotion label (**Neutral**, **Fearful**, **Dissatisfied**, **Apologetic**, **Abusive**, **Excited**, **Satisfied**), as well as **Macro** and **Weighted** F1 (excluding neutral). \* indicates statistically significant difference with  $p < 0.05$  between the best and the second best values in each column. For detailed results, please refer to Appendix F.1.

keep the original split of MultiWOZ and further divide DialMAGE with a ratio of 8:1:1, leading to 9,234, 1,100, and 1,100 dialogues in each set. We run each task on 5 different seeds and report the average performance.

### 5.1.3. Results and Discussion

**Recognition on emotion classes.** Table 5 summarises the performance of baseline models. Since almost 70% of the annotations are *neutral*, we exclude it when calculating average F1 scores. In general, models that take into account context information perform better on the full EmoWOZ. This shows the importance of context or dialogue-level features in emotion recognition in task-oriented dialogues. An exception is DialogueRNN with GloVe feature, which underperforms in EmoWOZ macro F1, likely due to the non-contextual embedding used. On the other hand, BERT scores very well on MultiWOZ dialogues but performs poorly on DialMAGE for both setups. This suggests that emotions in MultiWOZ are less context-dependent.

BERT, the only non-contextual model among our baselines, performs well for *apologetic*, *excited*, and *satisfied*, potentially due to the existence of distinguishable keywords associated with these emotions such as “thank you” for *satisfied* and “sorry” for *apologetic*. These emotion labels do not benefit much from context. In contrast, BERT produces a significantly worse F1 on *dissatisfied*, probably because users tend to express dissatisfaction more implicitly, for instance via repetition or correction, making dialogue-level features necessary.

Figure 2 shows two dialogues with implicit emotions and predictions made by respective baseline models. In example 1, the system gives the wrong time of arrival, eliciting mild annoyance from the user. BERT predicts *neutral* because in isolation, the utterance has no words

suggesting dissatisfaction. All other models correctly recognise *dissatisfied*, as they capture the misunderstanding occurs in previous dialogue turns. Example 2 presents a similar but more implicit case, where all models fail. This shows that EmoWOZ contains contextualised emotions that are more implicit and subtle, requiring more sophisticated features and models.

**Complementarity between MultiWOZ and DialMAGE.** Due to different linguistic features and emotion distributions in MultiWOZ and DialMAGE, one concern is that the models learn to predict emotion based on these statistical artifacts. According to Table 3, the most obvious difference is the average utterance length (5.8 in DialMAGE and 11.8 in MultiWOZ). A naive model may simply recognise the data source from word count and predict the most likely emotion from that source. Table 7 presents how ContextBERT trained on EmoWOZ predicts emotion in long DialMAGE and short MultiWOZ utterances. The emotion distribution in model prediction is vastly different from that in the complementing subset. Clearly, the model does not simply count words to decide on the underlying emotion.

	Dissatisfied	Satisfied
MultiWOZ Label	1.5%	24.0%
DialMAGE (#token>11.8) Label	28.8%	1.2%
DialMAGE (#token>11.8) Prediction	35.5%	1.5%
DialMAGE Label	39.3%	4.0%
MultiWOZ (#token<5.8) Label	1.2%	37.7%
MultiWOZ (#token<5.8) Prediction	3.0%	38.9%

Table 7: Emotion distribution in labels and ContextBERT prediction. See Appendix F.5 for full results.

Table 6 presents cross-data experiments with ContextBERT, examining how well the two subsets complement each other. Complementing DialMAGE with dialogues from MultiWOZ improves the macro F1 and the F1 score of *abusive* significantly. On the other

hand, while complementing MultiWOZ with DialMAGE leads to a slight improvement in the F1 score of *excited*, other F1 scores decrease to various extent.

**Recall and precision on satisfied and dissatisfied for task-oriented dialogue.** We further investigate the change in F1 of each emotion on MultiWOZ by looking at the change in recall and precision after complementing MultiWOZ with DialMAGE. We believe it is necessary to distinguish recall and precision, as for some emotions, one may be more important than the other. The relative importance of recall and precision for each emotion class depends on its implication to a task-oriented dialogue system and the consequence of false recognition. Most importantly for task-oriented dialogue system, a high recall of *dissatisfied* is desirable because the system should not miss any failure in dialogues. Failing to recognise dissatisfaction can trigger more anger from the user and therefore impair task completion (see Figure C.2). On the other hand, a high precision may be more desirable for all other emotions to ensure proper affective response from the system. When the relative importance of recall and precision of the emotion is taken into account, complementing MultiWOZ with DialMAGE is beneficial to *{dissatisfied}* for higher recall and *{fearful, excited, satisfied}* for higher precision, see Table 8. Detailed results can be found in Appendix F.4.

Metric	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.
Recall	-6.7**	+29.1**	-7.8	+8.0	+0.6	-0.4
Precision	+7.5*	-22.8**	-7.9	-11.3	+7.1**	+0.1

Table 8: Change in precision and recall for each emotion label (**F**earful, **D**issatisfied, **A**pologetic, **A**busive, **E**xcited, **S**atisfied) on MultiWOZ by ContextBERT, after adding DialMAGE to training. \*\* and \* indicate statistically significant changes with  $p < 0.05$  and  $p < 0.1$  respectively.

## 5.2. Emotions for Dialogue State Tracking

In task-oriented dialogues, dialogue state tracking (DST) aims to continuously track the user’s goal and intent as the dialogue progresses (Young et al., 2010). We hypothesise that the user emotion can help inform the system about their goal. To investigate this, we train a dialogue state tracker that incorporates an additional task to predict one of 7 emotional classes on MultiWOZ 2.1 (Eric et al., 2020). We utilise the *out-of-task training* approach and the available code presented in (Heck et al., 2020a). We follow the multitask learning (MTL) algorithm, where on each training step, the same model is trained on two different batches, one from the main task (DST) and one from the auxiliary task (emotion recognition). Since neutral emotion provides limited information on the user goal, we remove a half of the neutral utterances when performing MTL. We show that additional emotion labels can lead to a significant improvement ( $p < 0.02$ ) in the joint goal

accuracy (JGA) of DST (see Table 9).

Training tasks	JGA
Dialogue state tracking	53.7
Dialogue state tracking & emotion recognition	<b>54.7</b>

Table 9: JGA of DST on MultiWOZ.

## 6. Conclusion

In this work, we examined emotions and their expression in the context of task-oriented dialogues, where emotions are centred around a user goal. We used the OCC model as a starting point to derive a comprehensive annotation scheme beyond sentiment polarity for emotions in relation to user goals. We designed a set of 7 emotions that differ in terms of valence, conduct and elicitor to capture the cognitive context of emotions, while maintaining labeling simplicity.

With EmoWOZ, we present a publicly available, large-scale human-annotated emotion corpus consisting of Wizard-of-Oz style as well as dialogues with a machine-generated policy. Our intention with EmoWOZ is to overcome the lack of large emotion-labelled corpora to support research towards emotion-aware task-oriented dialogue systems, for dialogues closer to human-human interactions.

We apply various emotion recognition models to EmoWOZ and examined the effect of context for different emotions. In cross-dataset experiments we analysed the complementarity of WOZ-style data and machine-generated policy data. Our results show that recognising context-dependent and implicit emotions from task-oriented dialogues is a challenging task that will benefit from further research. EmoWOZ provides an ideal test bed for that. Lastly, we leveraged emotion recognition in the dialogue state tracking task to exemplify the utility of emotion labels in dialogue modeling. We hope this dataset can offer insights beyond the scope of emotion recognition and push the performance of downstream tasks in task-oriented dialogue modelling. In future work, we plan to investigate tailored models for emotion recognition in task-oriented dialogues that take advantage of high-level features such as dialogue acts or belief states. We are also interested in using emotion as a feedback signal within reinforcement learning policy optimisation.

## 7. Acknowledgements

S. Feng, N. Lubis, M. Heck, and C. van Niekerk are supported by funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research, while C. Geishauer and H-C. Lin are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018 804636). Computing resources were provided by Google Cloud.

## 8. Bibliographical References

- Bordes, A., Boureau, Y.-L., and Weston, J. (2017). Learning end-to-end goal-oriented dialog. In *ICLR*. OpenReview.net.
- Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July. Association for Computational Linguistics.
- Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October–November. Association for Computational Linguistics.
- Buechel, S. and Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.
- Carrión, Z. C. and López-Cózar, R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Commun.*, 50:416–433.
- Cercas Curry, A. and Rieser, V. (2018). #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA, June. Association for Computational Linguistics.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.
- Curry, A. C. and Rieser, V. (2019). A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366.
- Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural networks : the official journal of the International Neural Network Society*, 18 4:407–22.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, pages 169–200.
- Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May. European Language Resources Association.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.
- Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). COSMIC: CommonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online, November. Association for Computational Linguistics.
- Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.
- Gross, J. and Thompson, R. (2007). Emotion regulation: Conceptual foundations. *Handbook of Emotion Regulation*, pages 3–27, 01.
- He, W., Dai, Y., Zheng, Y., Wu, Y., Cao, Z., Liu, D., Jiang, P., Yang, M., Huang, F., Si, L., et al. (2022). Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Heck, M., Geishausser, C., Lin, H.-c., Lubis, N., Moresi, M., van Niekerk, C., and Gasic, M. (2020a). Out-of-task training for dialog state tracking models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6767–6774, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.
- Heck, M., van Niekerk, C., Lubis, N., Geishausser, C., Lin, H.-C., Moresi, M., and Gasic, M. (2020b). TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting, July. Association for Computational Linguistics.

- Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.
- Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, January.
- Kuppens, P., Van Mechelen, I., Smits, D. J., De Boeck, P., and Ceulemans, E. (2007). Individual differences in patterns of appraisal and anger experience. *Cognition and Emotion*, 21(4):689–713.
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.
- Li, Z., Kiseleva, J., and de Rijke, M. (2020). Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546, Online, November. Association for Computational Linguistics.
- Lin, W., Tseng, B.-H., and Byrne, B. (2021). Knowledge-aware graph-enhanced GPT-2 for dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Lubis, N., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (2015). Construction and analysis of social-affective interaction corpus in english and indonesian. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 202–206. IEEE.
- Lubis, N., Heck, M., Sakti, S., Yoshino, K., and Nakamura, S. (2017). Processing negative emotions through social communication: Multimodal database construction and analysis. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 79–85.
- Madotto, A., Lin, Z., Zhou, Z., Moon, S., Crook, P., Liu, B., Yu, Z., Cho, E., Fung, P., and Wang, Z. (2021). Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Mairesse, F. and Walker, M. (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic, June. Association for Computational Linguistics.
- Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825, Jul.
- Mayer, J. D., Caruso, D. R., and Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27(4):267–298.
- Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.
- Parrott, W. G. (2001). *Emotions in social psychology: essential readings*. Key readings in social psychology. Psychology Press, Philadelphia.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.
- Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA.
- Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July. Association for Computational Linguistics.
- Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S. Y. B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., Gelbukh, A., and Mihalcea, R. (2021). Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, Sep.
- Preoțiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (2016). Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California, June. Association for Computational Linguistics.
- Riccardi, G. and Hakkani-Tür, D. Z. (2005). Grounding emotions in human-machine conversational systems. In *INTETAIN*.
- Russell, J. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

- Saha, T., Saha, S., and Bhattacharyya, P. (2020). Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PLOS ONE*, 15(7):1–28, 07.
- Shi, W. and Yu, Z. (2018). Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia, July. Association for Computational Linguistics.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Ultes, S., Budzianowski, P., Casanueva, I., Mrksic, N., Rojas-Barahona, L. M., hao Su, P., Wen, T.-H., Gašić, M., and Young, S. J. (2017). Domain-independent user satisfaction reward estimation for dialogue policy learning. In *INTERSPEECH*.
- Vinyals, O. and Le, Q. (2015). A neural conversational model. *ICML Deep Learning Workshop, 2015*, 06.
- Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., Zhang, M., and Zhou, G. (2020). Sentiment classification in customer service dialogue with topic-aware multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184, Apr.
- Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., and Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.
- Young, S. (2002). Talking to machines (statistically speaking). In *Seventh International Conference on Spoken Language Processing*.
- Zahiri, S. and Choi, J. D. (2018). Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis, AFCON’18*, pages 44–51, New Orleans, LA.
- Zhou, X. and Wang, W. Y. (2018). MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia, July. Association for Computational Linguistics.
- Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI’18/IAAI’18/EAAI’18*. AAAI Press.
- Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., and Huang, M. (2020). ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In Asli Celikyilmaz et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 142–149. Association for Computational Linguistics.

## A. The OCC Model

Figure A.1 summarises definitions of emotion groups in the OCC model.

Elicitor	Aspects of events or agents	OCC Emotion		
Consequences of Events	Consequence for other	Desirable for other	happy-for resentment	
		Undesirable for other	gloating pity	
	Consequence for self	prospects relevant	confirmed	satisfaction fears-confirmed
		prospects irrelevant	disconfirmed	relief disappointment
				joy distress
	Actions of Agents	Consequence for self, prospect irrelevant, and related to actions of agents.	self agent	gratification remorse
other agent			gratitude anger	
self agent		pride shame		
other agent		admiration reproach		
Aspects of Objects		love hate		

Figure A.1: The OCC Model

## B. Amazon Mechanical Turk Set-up

### B.1. Qualification Test

Figure B.1 illustrates one example from our qualification test. Hints are provided for difficult questions containing implicit emotions as shown in the example.

In Question 11 - 12, the user repeatedly ask about something similar. Please try to think why the user repeatedly ask about something similar and infer the user's emotion from the context.

Question 11

(User: I need a taxi from the hotel to the museum after 23:45)

(Operator: Do you want the hotel reservations to begin on monday ? ...)

(User: We're talking about a taxi now)

(Operator: You would love broughton house gallery...)

User: Taxi.

- Neutral** (The user does not show obvious emotions when user is, e.g., asking for information, describing searching criteria, and saying byes. You (as the operator) may just want to respond the user.)
- Sad/fear** (Negative emotions caused by events or facts rather than the operator. E.g. user encountered in injury/accident/robbery, booking not available. You (as the operator) may feel empathetic and want to comfort the user.)
- Disliking/dissatisfied** (Negative emotions caused by the operator during the dialogue. E.g. user not happy with the operator's mistake or suggestion. You (as the operator) may feel apologetic for mistakes made.)
- Apologetic** (E.g. user apologised for his/her mistakes, changing search criteria, causing inconvenience or confusion to the operator. You (as the operator) may want to relieve the user by saying "no worries")
- Angry/abusive** (The user is extreme angry and even insulting the operator. You feel offended if you were the operator.)
- Anticipating/happy/excitement** (Positive emotions caused by events or facts. E.g. user looking forward to or excited about a holiday, birthday, anniversary, tour attraction, etc. You (as the operator) may feel happy for the user.)
- Liking/satisfied/appreciative/grateful** (Positive emotions caused by the dialogue. E.g. user happy with the operator's help or suggestion. You (as the operator) feel encouraged and know that you are doing the right job.)

Figure B.1: One of fifteen questions in our qualification test

### B.2. Main Task Page

Figure B.2 shows the task page for workers. Before arriving at this page, they will be prompted with a consent form and a message asking if they would like to go through a tutorial.

Instructions Please select the group of emotions that best describes the highlighted sentence.

Dialogue	Your Work
Please label the highlighted dialogue below. (Progress 1/5)  <b>User: am looking for a place to to stay that has cheap price range it should be in a type of hotel</b>	<ul style="list-style-type: none"><li><input type="radio"/> <b>Neutral</b> The user does not show obvious emotions when user is, e.g., asking for information, describing searching criteria, and saying byes. You (as the operator) may just want to respond the user.</li><li><input type="radio"/> <b>Sad/fear</b> Negative emotions caused by events or facts rather than the operator. E.g. user encountered in injury/accident/robbery, booking not available. You (as the operator) may feel empathetic and want to comfort the user.</li><li><input type="radio"/> <b>Disliking/dissatisfied</b> Negative emotions caused by the operator during the dialogue. E.g. user not happy with the operator's mistake or suggestion. You (as the operator) may feel apologetic for mistakes made.</li><li><input type="radio"/> <b>Apologetic</b> E.g. user apologised for his/her mistakes, changing search criteria, causing inconvenience or confusion to the operator. You (as the operator) may want to relieve the user by saying "no worries".</li><li><input type="radio"/> <b>Angry/abusive</b> The user is extreme angry and even insulting the operator. You feel offended if you were the operator.</li><li><input type="radio"/> <b>Anticipating/happy/excitement</b> Positive emotions caused by events or facts. E.g. user looking forward to or excited about a holiday, birthday, anniversary, tour attraction, etc. You (as the operator) may feel happy for the user.</li><li><input type="radio"/> <b>Liking/satisfied/appreciative/grateful</b> Positive emotions caused by the dialogue. E.g. user happy with the operator's help or suggestion. You (as the operator) feel encouraged and know that you are doing the right job.</li></ul>

prev next submit

Disliking/dissatisfied Add examples()

Example 1  
(Operator: i have booked you clare hall)  
Users: what? that's not what i wanted.  
Example 2 (Implicit annoyance)  
(The user was asking for a restaurant in the south)  
(Operator: What attraction were you thinking of? ...)  
(User: Restaurants south)  
(Operator: What cuisine are you interested in ?)  
User: Restaurant in the south please.

Figure B.2: Amazon Mechanical Turk main task page

## C. Dialogue Examples

Figure C.1 shows examples of how emotions are expressed by the user in EmoWOZ. Figure C.2 shows examples of annotated dialogues.

Emotion	Example
Neutral	Operator: What sort of food would you like it to be?) User: You choose. Book me a table for 6 people at 12:00 on Thursday.
Fearful, sad, disappointed	(Start of the dialogue) User: Can you help me please? I just got robbed and I need help!! Help!
Dissatisfied, disliking	(Explicit expression) Operator: i have booked you clare hall User: what? that's not what i wanted.
	(Implicit expression) User: I need a taxi from the hotel to the museum after 23:45 Operator: Do you want the hotel reservations to begin on Monday?... User: We're talking about a taxi now Operator: You would love broughton house gallery User: Taxi.
Apologetic	Operator: I'm afraid that's all we've got. User: I'm sorry I may have confused you. I would want to go from Petersborough to Cambridge on Saturday and arrive by 13:00, so anything that will get me there before that.
Abusive	Operator: (talking about something irrelevant) User: **** **, would ya stupid thing
Excited, happy, anticipating	Operator: There are 3 restaurants on the west side that serve British food. Do you prefer moderate or expensive priced? User: Expensive, please, we're splashing out for my mother's birthday
Satisfied, liking, appreciative, grateful	Operator: Any time. I was happy to help. User: You are doing a great job! That's all I needed

Figure C.1: Example for each emotion label

Annotation Example 1 (MultiWOZ)	
Operator:	... Is there anything else I can help with today?
User:	Yes I would like to visit some architecture landmarks in the centre of the city. What would you recommend? [Excited]
Operator:	yes, there are four churches and a school in that area.
User:	Could you suggest one to me? I will also need to know the address and phone number, as well as the entrance fee. Thanks. [neutral]
Operator:	all saints church is free admission and is located at jesus lane. The phone number is 01223452587.
User:	Thanks so much! That is all I need help with today! [appreciative]
Operator:	You're quite welcome! Have a lovely day, goodbye!
Annotation Example 2 (DialIMAGE)	
Operator:	Where is your destination ?
User:	liverpool [neutral]
Operator:	The arrival time is 05:51 . The train is for sunday you are all set. The train will be departing from cambridge . The booking is for arriving in london kings cross .
User:	thats great...to bad the city is completely wrong but ok [dissatisfied]
	... ..
User:	yes book a return on Monday [neutral]
Operator:	Did you need any further assistance today ?
User:	yes please book a return on Monday [dissatisfied]
Operator:	Is there anything else I can help you with ?
User:	BOOK A RETURN TRAIN ON MONDAY [abusive]

Figure C.2: Annotation examples from EmoWOZ

## D. Annotator Agreement in EmoWOZ

Source	Fleiss' Kappa	#NA	#PA	#FA
MultiWOZ	0.611	1,016	17,270	53,238
DialMAGE	0.465	259	4,773	7,061
EmoWOZ (overall)	0.602	1,275	22,043	60,299

Table D1: Inter-annotator agreement and agreement count of EmoWOZ and its subsets. NA means no agreement—three annotators annotate with three different emotions. PA means partial agreement—only two annotators annotate with the same emotion. FA means full agreement—three annotators annotate with the same emotion.

## E. Hyperparameters for Model Training

Model	Optimiser	Learning Rate	L2 Regulariser Weight	Training Epochs
BERT	Adam	2e-5	0	10
ContextBERT	Adam	2e-5	0	10
DialogueRNN(GloVe)	Adam	1e-4	1e-5	60
DialogueRNN(BERT)	Adam	1e-4	1e-4	60
COSMIC	Adam	1e-4	3e-4	20

Table E1: Hyperparameters for model training

## F. Detailed Cross-dataset Experiment Results

### F.1. Emotion Classification (7 classes)

Model	Set-up	F1 for each Emotion Label							Average F1 w/o Neutral			Average F1 w Neutral		
		Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied	Micro	Macro	Weighted	Micro	Macro	Weighted
BERT	D → D	59.75	0	<b>50.34</b>	0	12.99	61.42	<b>72.43</b>	<b>52.50</b>	32.86	<b>51.45</b>	59.08	36.70	55.94
	M → D	<b>71.57</b>	<b>11.67</b>	1.36	<b>100</b>	6.15	<b>64.30</b>	68.85	16.97	42.05	9.36	56.94	46.27	43.02
	E → D	69.94	0	41.43	60.0	<b>29.41</b>	56.01	69.13	45.47	<b>42.66</b>	43.82	<b>61.09</b>	<b>46.56</b>	<b>57.95</b>
	D → M	71.09	0	6.02	0	11.11	15.60	88.07	62.63	20.13	77.16	70.58	27.41	72.77
	M → M	<b>95.34</b>	<b>43.00</b>	<b>40.87</b>	<b>73.03</b>	19.05	<b>40.45</b>	<b>90.39</b>	<b>85.19</b>	<b>51.13</b>	<b>84.82</b>	<b>92.57</b>	<b>57.45</b>	<b>92.43</b>
	E → M	92.67	41.43	27.76	70.35	<b>21.43</b>	39.98	89.44	79.79	48.4	83.19	88.88	54.72	90.05
E → E	89.75	36.17	35.10	70.38	27.50	42.89	88.79	73.67	50.14	73.55	84.82	55.80	84.83	
ContextBERT	D → D	80.16	0	<b>75.69</b>	0	5	58.58	<b>71.69</b>	<b>73.91</b>	35.16	72.85	77.19	41.59	76.81
	M → D	72.11	<b>11.67</b>	7.73	<b>43.71</b>	11.87	<b>60.07</b>	66.29	21.29	33.56	14.49	57.80	39.06	45.67
	E → D	<b>81.58</b>	5.00	75.46	40.00	<b>52.81</b>	57.31	69.23	73.71	<b>49.97</b>	<b>73.49</b>	<b>77.89</b>	<b>54.48</b>	<b>77.87</b>
	D → M	89.37	0	11.18	0	0	13.86	77.07	59.43	17.02	67.81	80.44	27.35	83.40
	M → M	<b>95.09</b>	<b>35.71</b>	<b>36.35</b>	<b>70.34</b>	<b>19.44</b>	34.05	<b>90.01</b>	<b>84.36</b>	<b>47.65</b>	<b>83.87</b>	<b>92.14</b>	<b>54.43</b>	<b>91.98</b>
	E → M	93.45	33.70	30.39	62.42	17.27	<b>37.06</b>	89.75	80.44	45.10	83.14	89.65	52.00	90.60
E → E	92.10	30.08	61.69	62.36	41.73	40.83	89.14	78.99	54.30	79.67	87.93	59.70	88.33	
DialogueRNN (GloVe)	D → D	40.13	0	<b>64.01</b>	0	0	52.05	65.59	<b>62.56</b>	30.28	<b>62.03</b>	<b>54.88</b>	31.68	<b>50.18</b>
	M → D	<b>67.00</b>	0	22.91	<b>100</b>	0	54.45	55.96	31.03	38.89	26.23	54.07	<b>42.90</b>	48.30
	E → D	23.83	0	61.75	60	0	<b>63.72</b>	<b>73.57</b>	62.06	<b>43.17</b>	61.23	50.24	40.41	40.99
	D → M	59.78	0	5.34	0	0	13.80	85.57	50.51	17.45	74.87	55.46	23.50	63.96
	M → M	87.25	<b>21.57</b>	<b>21.53</b>	52.16	0	26.21	85.51	72.78	<b>34.50</b>	78.12	82.21	<b>42.03</b>	84.72
	E → M	<b>88.24</b>	13.59	18.67	<b>57.56</b>	0	<b>27.92</b>	<b>86.73</b>	<b>74.04</b>	34.08	<b>79.22</b>	<b>83.41</b>	41.81	<b>85.74</b>
E → E	83.46	12.71	51.38	57.67	0	32.75	86.35	70.93	40.14	74.56	78.56	46.33	80.76	
DialogueRNN (BERT)	D → D	65.24	0	58.24	0	27.69	54.51	<b>68.35</b>	58.45	34.80	57.97	<b>61.95</b>	39.15	<b>61.90</b>
	M → D	<b>66.63</b>	0	4.24	43.52	2.86	41.48	53.87	17.33	24.33	9.64	50.73	30.37	40.48
	E → D	49.81	0	<b>61.01</b>	<b>91.67</b>	<b>28.14</b>	<b>60.92</b>	66.70	<b>60.95</b>	<b>51.41</b>	<b>60.56</b>	56.58	<b>51.18</b>	54.74
	D → M	85.48	0	8.17	0	6.71	20.48	87.46	65.74	20.47	76.91	78.71	29.76	83.11
	M → M	<b>92.11</b>	34.83	<b>34.49</b>	58.59	0	26.32	87.48	<b>79.18</b>	40.28	80.84	<b>88.13</b>	47.69	<b>88.99</b>
	E → M	90.54	<b>47.12</b>	18.08	<b>71.22</b>	<b>15.48</b>	<b>33.92</b>	<b>88.28</b>	76.26	<b>45.68</b>	<b>81.52</b>	86.10	<b>52.09</b>	88.04
E → E	86.85	41.32	47.51	71.48	25.56	39.42	87.58	72.48	52.15	75.50	81.78	57.10	83.41	
COSMIC	D → D	69.34	0	59.68	0	0	64.30	<b>72.25</b>	60.31	32.71	59.25	<b>65.07</b>	37.94	<b>64.71</b>
	M → D	<b>71.56</b>	33.33	2.67	100	15.38	67.04	70.80	19.98	48.21	11.03	57.16	51.54	43.79
	E → D	66.59	0	<b>61.47</b>	<b>100</b>	<b>43.71</b>	<b>69.74</b>	68.19	<b>62.09</b>	<b>57.18</b>	<b>61.67</b>	64.47	<b>58.53</b>	64.33
	D → M	86.68	0	8.78	0	0	20.91	88.90	67.85	19.77	78.19	80.32	29.32	84.33
	M → M	<b>94.86</b>	50	<b>40.97</b>	67.12	0	41.77	<b>89.93</b>	<b>84.22</b>	<b>48.30</b>	<b>84.27</b>	<b>91.81</b>	<b>54.95</b>	<b>91.93</b>
	E → M	92.61	<b>58.18</b>	24.68	<b>70.52</b>	0	<b>37.92</b>	89.10	79.84	46.73	82.74	88.81	53.29	89.88
E → E	89.80	51.98	50.69	70.93	31.62	44.42	88.42	75.89	56.34	77.09	85.26	61.12	85.94	

Table F1: Performance of baseline models on emotion classification including cross-dataset experiments. For cross-dataset experiments, the “X → Y”s in the ‘Set-up’ column represents the training and evaluation set-up, where X is the training set and Y is the test set. E stands for EmoWOZ, M stands for MultiWOZ, and D stands for DialMAGE. M → D, for example, means to train on MultiWOZ and test on DialMAGE. Extreme values for “Apologetic” and “Abusive” in DialMAGE (“\* → D”s) are caused by their rarity in the test set (1 and 5 occurrences respectively).

### F.2. Confusion Matrix of ContextBERT

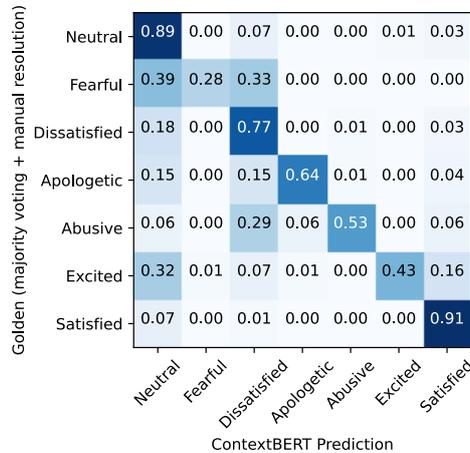


Figure F.1: Confusion matrix between Golden Labels and (the best) ContextBERT Prediction.

### F.3. Sentiment Classification (3 classes)

Model	Feature	Ctx.	F1 of Each Sentiment in EmoWOZ			Average F1 w/o Neutral					
			Neutral	Negative	Positive	EmoWOZ		MultiWOZ		DialMAGE	
						Macro	Weighted	Macro	Weighted	Macro	Weighted
BERT	BERT	No	89.9	42.0	88.3	65.1	75.6	67.0	84.6	54.8	44.7
ContextBERT	BERT	Yes	<b>92.7</b>	<b>68.0</b>	87.5	<b>77.8</b>	<b>82.2</b>	<b>70.2</b>	<b>84.7</b>	<b>71.2</b>	<b>75.3</b>
DialogueRNN	GloVe	Yes	86.8	61.0	83.7	72.4	77.5	66.5	80.6	69.0	66.7
DialogueRNN	BERT	Yes	83.3	47.2	87.4	67.3	76.4	57.0	81.8	67.2	66.1
COSMIC	BERT+COMET	Yes	89.0	48.2	<b>88.5</b>	68.3	77.5	63.0	84.0	63.7	59.1

Table F2: Summarised performance of baseline models on sentiment classification. ‘‘Ctx.’’ stands for ‘‘context’’.

Model	Set-up	F1 for each Sentiment Label			Average F1 w/o Neutral			Average F1 w Neutral		
		Neutral	Negative	Positive	Micro	Macro	Weighted	Micro	Macro	Weighted
BERT	D → D	57.61	<b>61.32</b>	<b>72.48</b>	<b>62.45</b>	<b>66.90</b>	<b>62.61</b>	60.54	<b>63.80</b>	<b>59.91</b>
	M → D	<b>71.70</b>	3.80	68.47	17.78	36.13	11.27	57.39	47.99	43.93
	E → D	70.01	41.73	67.84	46.16	54.78	44.74	<b>61.68</b>	59.86	58.39
	D → M	71.97	7.77	84.71	56.23	46.24	76.94	65.82	54.82	73.35
	M → M	<b>95.43</b>	<b>56.68</b>	<b>90.05</b>	<b>87.00</b>	<b>73.37</b>	<b>86.69</b>	<b>93.11</b>	<b>80.72</b>	<b>92.99</b>
	E → M	92.90	44.99	89.08	81.99	67.04	84.63	89.66	75.66	90.60
	E → E	89.93	41.96	88.27	75.74	65.11	75.60	85.57	73.39	85.56
ContextBERT	D → D	80.67	<b>76.74</b>	<b>70.81</b>	<b>76.08</b>	<b>73.77</b>	<b>76.05</b>	78.53	<b>76.07</b>	78.55
	M → D	71.69	6.65	64.22	20.02	35.43	13.30	57.27	47.52	44.85
	E → D	<b>82.91</b>	76.60	65.71	75.25	71.15	75.34	<b>79.48</b>	75.07	<b>79.43</b>
	D → M	89.04	17.89	69.53	54.85	43.71	64.32	79.01	58.82	82.16
	M → M	<b>95.38</b>	<b>55.09</b>	<b>89.89</b>	<b>86.81</b>	<b>72.49</b>	<b>86.38</b>	<b>93.03</b>	<b>80.12</b>	<b>92.88</b>
	E → M	93.98	51.94	88.38	83.75	70.16	84.70	91.05	78.10	91.40
	E → E	92.69	67.98	87.55	81.95	77.76	82.19	89.36	82.74	89.49
DialogueRNN (GloVe)	D → D	58.50	<b>69.49</b>	67.55	<b>69.21</b>	68.52	69.26	<b>64.98</b>	<b>65.18</b>	<b>63.45</b>
	M → D	<b>68.70</b>	20.44	52.21	29.70	36.32	24.11	55.21	47.12	48.21
	E → D	42.57	66.00	<b>71.96</b>	66.56	<b>68.98</b>	<b>66.68</b>	58.16	60.17	53.65
	D → M	77.27	10.88	<b>84.93</b>	62.22	47.90	77.46	71.38	57.69	77.32
	M → M	<b>91.06</b>	<b>50.36</b>	84.13	<b>80.22</b>	<b>67.25</b>	<b>80.72</b>	<b>87.57</b>	<b>75.19</b>	<b>88.18</b>
	E → M	90.71	48.85	84.21	79.77	66.53	80.64	87.14	74.59	87.91
	E → E	86.75	61.03	83.74	76.42	72.39	77.53	82.91	77.18	83.94
DialogueRNN (BERT)	D → D	47.53	65.27	<b>69.53</b>	65.74	<b>67.40</b>	65.76	59.05	60.78	55.91
	M → D	<b>71.10</b>	7.04	64.41	20.25	35.72	13.67	56.77	47.52	44.70
	E → D	50.18	<b>65.81</b>	68.63	<b>66.10</b>	67.22	<b>66.14</b>	<b>60.04</b>	<b>61.54</b>	<b>57.52</b>
	D → M	59.37	8.08	86.36	51.21	47.22	78.46	55.78	51.27	64.68
	M → M	<b>93.75</b>	<b>52.28</b>	<b>88.13</b>	<b>84.27</b>	<b>70.20</b>	<b>84.52</b>	<b>90.94</b>	<b>78.05</b>	<b>91.18</b>
	E → M	86.61	25.98	88.11	73.03	57.04	81.84	81.92	66.90	85.29
	E → E	83.30	47.16	87.36	71.40	67.26	76.36	78.72	72.61	81.19
COSMIC	D → D	63.51	<b>68.32</b>	<b>70.52</b>	<b>68.58</b>	<b>69.42</b>	<b>68.57</b>	<b>66.45</b>	<b>67.45</b>	<b>65.83</b>
	M → D	<b>71.69</b>	5.22	67.57	18.79	36.39	12.42	57.32	48.16	44.45
	E → D	69.28	57.68	69.66	59.30	63.67	59.06	64.80	65.54	64.58
	D → M	80.31	12.43	86.38	62.62	49.41	78.92	73.66	59.71	79.92
	M → M	<b>95.01</b>	<b>57.85</b>	<b>89.48</b>	<b>86.41</b>	<b>73.67</b>	<b>86.29</b>	<b>92.58</b>	<b>80.78</b>	<b>92.58</b>
	E → M	91.63	36.68	89.32	79.78	63.00	84.01	87.99	72.54	89.51
	E → E	88.97	48.15	88.55	75.66	68.35	77.5	84.6	75.22	85.47

Table F3: Detailed results of baseline models on sentiment classification including cross-dataset experiments. For cross-dataset experiments, the ‘‘X → X’’s in the ‘Set-up’ column represents the training and evaluation set-up. E stands for EmoWOZ, M stands for MultiWOZ, and D stands for DialMAGE. M → D, for example, means to train on MultiWOZ and test on DialMAGE.

#### F.4. Change in precision and recall on MultiWOZ after Complementing MultiWOZ with DialMAGE in Training

		Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied
ContextBERT	Recall	<b>95.3</b> → <b>91.5</b>	<b>34.7</b> → <b>28.0</b>	<b>31.4</b> → <b>60.4</b>	69.7 → 61.9	16.0 → 24.0	33.5 → 34.1	90.4 → 90.0
	Precision	<b>94.9</b> → <b>95.5</b>	<b>37.3</b> → <b>44.8</b>	<b>43.7</b> → <b>20.9</b>	71.4 → 63.5	25.0 → 13.7	<b>35.5</b> → <b>42.6</b>	89.5 → 89.6
	F1	<b>95.1</b> → <b>93.5</b>	35.7 → 33.7	36.4 → 30.4	<b>70.3</b> → <b>62.4</b>	19.4 → 17.3	34.0 → 37.1	90.0 → 89.7

Table F4: Precision, recall and F1 score of ContextBERT for all emotions when trained on MultiWOZ and EmoWOZ respectively, and tested on MultiWOZ.  $A \rightarrow B$  represents how the value change after complementing MultiWOZ with DialMAGE in training.  $A$  is the value when trained on MultiWOZ and  $B$  is the value when trained on EmoWOZ. Values with statistical significance ( $p < 0.1$ ) are bolded and colored where red indicates a drop and green indicates an improvement. For recognising user emotions in task-oriented dialogues, a high precision is more desirable for *neutral*, *fearful*, *apologetic*, *abusive*, *excited*, and *satisfied* where as a high recall is more desirable for *dissatisfied*.

#### F.5. Emotion Distribution in Model Predictions

Test Set	Model	Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied
MultiWOZ Label		72.31	0.2	1.47	0.98	0.07	1.0	23.97
DialMAGE (#token>11.8) Label		61.96	0.61	28.83	0.61	0.0	6.75	1.23
DialMAGE (#token > 11.8) Prediction	BERT	89.57	0.0	2.45	0.0	0.0	6.75	1.23
	ContextBERT	58.13	0.0	35.47	0.0	0.0	4.93	1.48
	DialogueRNN-GloVe	6.52	0.0	74.97	1.28	0.0	6.05	11.18
	DialogueRNN-BERT	55.65	0.12	23.05	8.61	0.0	5.01	7.57
	COSMIC	61.93	0.23	18.63	7.57	0.0	3.84	7.8
DialMAGE Label		54.12	0.24	39.3	0.08	0.95	1.35	3.96
MultiWOZ (#token<5.8) Label		60.76	0.0	1.21	0.0	0.0	0.3	37.73
MultiWOZ (#token < 5.8) Prediction	BERT	60.91	0.15	0.45	0.3	0.0	0.45	37.73
	ContextBERT	57.43	0.0	2.97	0.0	0.0	0.66	38.94
	DialogueRNN-GloVe	46.54	0.0	2.1	0.7	0.0	0.88	49.78
	DialogueRNN-BERT	46.28	0.0	8.94	0.26	0.09	1.67	42.77
	COSMIC	49.43	0.09	5.0	0.18	0.09	1.58	43.65

Table F5: Emotion distribution in model predictions (trained on EmoWOZ).



## Chapter 6

# From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue

This chapter summarises our work on developing a framework that leverages characteristics of task-oriented dialogues (ToDs) and cross-domain resources to further enhance emotion recognition in ToDs and gives a verbatim copy of our paper (Feng et al., 2023):

Shutong Feng et al. (Sept. 2023). “From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue”. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Svetlana Stoyanchev et al. Prague, Czechia: Association for Computational Linguistics, pp. 85–103. DOI: 10.18653/v1/2023.sigdial-1.8. URL: <https://aclanthology.org/2023.sigdial-1.8/>

## 6.1 Summary

From our previous work (Chapter 5), we identify several challenges in emotion recognition learning in ToDs as a result of inherent data imbalance and more nuanced emotion expressions. Unlike open-domain chit-chat, ToDs require emotion modelling that accounts for goal-driven structure and context. We therefore propose a framework to adapt emotion recognition models to ToDs, where emotions are closely tied to task success and system performance.

Our framework consists of approaches from three aspects: (1) data, by proposing emotion-focused augmentation strategies for the EmoWOZ dataset; (2) features, by incorporating dialogue state and task progress as input features, coupled with a sentiment-optimised feature extractor; and (3) loss design, introducing an emotion distance-weighted loss and a multi-task learning loss to improve emotionally relevant predictions.

Our experiments show significant improvement in user emotion recognition under the task-oriented context after applying our framework to existing models built for chit-chat. We further explore the generalisation of the best resulting model to a related task, in particular, user satisfaction prediction. We observe a strong zero-shot performance in this task. This work highlights the importance of task-specific modelling in building emotion-aware dialogue systems.

## 6.2 Personal Contributions

The implementation, technical results, and writing are my own work. Milica Gasić supervised the project. Nurul Lubis co-supervised the project. Co-authors assisted in writing and proofreading.

# From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishaus, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gašić

Heinrich Heine University Düsseldorf, Germany

{fengs, lubis, ruppik, geishaus, heckmi, linh, niekerk, revuk100, gasic}@hhu.de

## Abstract

Emotion recognition in conversations (ERC) is a crucial task for building human-like conversational agents. While substantial efforts have been devoted to ERC for chit-chat dialogues, the task-oriented counterpart is largely left unattended. Directly applying chit-chat ERC models to task-oriented dialogues (ToDs) results in suboptimal performance as these models overlook key features such as the correlation between emotions and task completion in ToDs. In this paper, we propose a framework that turns a chit-chat ERC model into a task-oriented one, addressing three critical aspects: data, features and objective. First, we devise two ways of augmenting rare emotions to improve ERC performance. Second, we use dialogue states as auxiliary features to incorporate key information from the goal of the user. Lastly, we leverage a multi-aspect emotion definition in ToDs to devise a multi-task learning objective and a novel emotion-distance weighted loss function. Our framework yields significant improvements for a range of chit-chat ERC models on EmoWOZ, a large-scale dataset for user emotion in ToDs. We further investigate the generalisability of the best resulting model to predict user satisfaction in different ToD datasets. A comparison with supervised baselines shows a strong zero-shot capability, highlighting the potential usage of our framework in wider scenarios.

## 1 Introduction

Emotion recognition in conversations (ERC) is a crucial task in conversational artificial intelligence research because it lays the foundation for affective abilities in computers such as empathetic response generation (Picard, 1997). Over years, it has shown values in downstream applications such as opinion mining (Colneric and Demšar, 2020) and human-like dialogue modelling (Zhou et al., 2018).

Dialogue systems can be broadly categorised into two categories: (1) chit-chat or open-domain

```
A: We have a holiday next week, don't we ?
B: Yes, on Monday .
A: What're you going to do ?
B: I'm probably going to spend the day looking at cars.
...
```

(a) Chit-chat dialogue from Li et al. (2017)

```
U: I am planning a vacation and really could use some
  help finding a good place to stay in town. I've
  never been to Cambridge before.
S: which side of town do you prefer and what is the
  price range?
U: It doesn't matter. What do you recommend?
S: alexander bed and breakfast is a guest house in the
  centre area. Would you like to book a room?
...
```

(b) Task-oriented dialogue from Budzianowski et al. (2018)

Figure 1: Comparison of dialogues about holiday in chit-chat dialogues and task-oriented dialogues.

systems and (2) task-oriented dialogue (ToD) systems. Chit-chat systems are set up to mimic human behaviours in a conversation (Jurafsky and Martin, 2009). There are no particular goals associated with the dialogue and the system aims to keep the user engaged with natural and coherent responses. On the other hand, ToD systems are concerned with fulfilling user goals, such as information retrieval for hotel booking (Young, 2002).

Recently, the difference between chit-chat and ToD systems have been blurred by the utilisation of pre-trained language models as back-bone to both types of systems. However, emotions in ToDs and chit-chat dialogues play different roles and are therefore expressed differently (Feng et al., 2022). This highlights the need for dedicated emotion modelling methods for each system.

As illustrated in Figure 1, in chit-chat dialogues, speakers make use of emotions to facilitate communication by, for example, raising empathy as a result of emotion-eliciting situations or topics. On the other hand, emotions in ToDs are centred around the user's goal, and therefore emotion cues lie in both the user's wording and the task performance.

While many large-scale corpora for emotions in chit-chat dialogues exist (Busso et al., 2008; McKeown et al., 2012; Lubis et al., 2015; Li et al., 2017; Zahiri and Choi, 2018), there are considerably fewer resources for emotions in ToDs. EmoWOZ, which evolved from MultiWOZ, a widely used ToD dataset, is one notable exception (Feng et al., 2022). It contains a novel emotion description that is designed for ToDs and inspired by the Ortony-Clore-Collins (OCC) model (Ortony et al., 1988). Emotion is described in terms of three aspects: **valenced** (positive or negative) reactions towards **elicitors** (operator, user, or event) in a certain **conduct** (polite or impolite). However, due to the nature of ToDs, the occurrence of some emotions (e.g. users expressing feelings about their situations) are very rare, leading to a class imbalance in the corpus.

Similarly, advancements on the ERC task are mainly focused on chit-chat dialogues, involving an array of diverse factors from speaker personality (Majumder et al., 2019) to commonsense knowledge (Ghosal et al., 2020). Nevertheless, since these models are designed for chit-chat dialogues, they overlook how emotions are triggered and expressed with respect to goal completion in task-oriented context. The work of Devillers et al. (2003) is among one of the earliest and very few to address emotion detection in ToDs but uses generic unigram models instead of dedicated approaches.

In this work, we tackle critical steps of ERC in ToDs from three angles: the data, the features, and the learning objective. In particular,

**Data:** we address the poor ERC performance of particularly rare emotions in ToDs via two strategies of data augmentation (DA),

**Features:** we leverage dialogue state information and sentiment-aware textual features,

**Objective:** we exploit the three aspects of emotions, namely valence, elicitor, and conduct, in two ways: as a multi-task learning (MTL) objective and to define a novel emotion-distance-weighted loss (*EmoDistLoss*).

To the best of our knowledge, our work is the first to provide dedicated methods for emotion recognition in ToDs. Our experiments and analyses show that our framework leads to significant improvements for a range of chit-chat ERC models when evaluated on EmoWOZ.

We further investigate the generalisability of the best resulting model to predict user satisfaction in

various ToD datasets under zero-shot transfer. Our model achieves comparable results as supervised baselines, demonstrating strong zero-shot capability and potential to be applied in wider scenarios.

## 2 Related Work

### 2.1 ERC Datasets

Early work on ERC relied on small scale datasets (Busso et al., 2008; McKeown et al., 2012; Lubis et al., 2015). More recently, a few large-scale datasets have been made available to the research community. They contain dialogues from emotion-rich and spontaneous scenarios such as daily communications (Li et al., 2017) and situation comedies (Zahiri and Choi, 2018).

For ToDs, the majority of available datasets address only one particular aspect of emotions such as sentiment polarity (Saha et al., 2020; Shi and Yu, 2018), user satisfaction (Schmitt et al., 2012; Sun et al., 2021), and politeness (Hu et al., 2022; Mishra et al., 2023). For more fine-grained emotions, Singh et al. (2022) constructed EmoInHindi for emotion category and intensity recognition in mental health and legal counselling dialogues in Hindi, and Feng et al. (2022) released EmoWOZ, which concerns user emotions in human-human and human-machine in information-seeking dialogues. Among these datasets, EmoWOZ has the largest scale, accompanied with a label set tailored to the task-oriented scenario.

### 2.2 Data Augmentation (DA)

DA is an effective approach to improve model performance by improving data diversity without explicitly collecting more data. While textual DA can be performed in the feature space via interpolation and sampling (Kumar et al., 2019), it is commonly performed in the data space for controllability. Rule-based methods involve operations such as insertion and substitution (Wei and Zou, 2019). While they are easy to implement, the diversity in augmented samples depends on the complexity of the rules. On the contrary, model-based methods are more scalable. These typically include the use of language models (Jiao et al., 2020), translation models (Xie et al., 2020a), and paraphrasing methods (Hou et al., 2018).

Additional training samples can also be obtained from unlabelled data via weak supervision (Ratner et al., 2017). To generate the automatic labels, a single model or an ensemble of models may

be used. This method can be interpreted as self-augmentation (Xu et al., 2022), self-training (Xie et al., 2020b), or distillation (Radosavovic et al., 2017).

DA has also been also deployed in ToD modelling. Hou et al. (2018) generated samples by paraphrasing delexicalised utterances. Gritta et al. (2021) conceptualised ToDs into transitional graphs and generate new dialogue paths by sampling. Heck et al. (2022) proposed a weak supervision framework to address the lack of fine-grained span labels for dialogue state tracking. DA for emotions in ToDs requires careful considerations to avoid emotion mismatch and is not yet explored.

### 2.3 ERC Models and Features

Text-based ERC is in essence a text classification problem with an emphasis on contextual modelling. Poria et al. (2017) proposed a recurrent neural network (RNN) for multimodal ERC. The follow-up work of Majumder et al. (2019) considered speaker-specific context. ERC performance has been continuously improved by techniques such as incorporating external knowledge (Ghosal et al., 2020) and contrastive learning (Song et al., 2022).

**Sentiment-aware Embeddings** Word-vector embeddings tailored for a particular natural language processing task can effectively improve the performance for that task (Naseem et al., 2021). In a similar vein, Tang et al. (2014) incorporated sentiment classification objectives in the training of the word embedding model of Collobert and Weston (2008) specifically for sentiment analysis. Yu et al. (2017) refined static word embeddings with the aid of a sentiment lexicon. Later, many sentiment-aware variants of pre-trained language models were obtained by incorporating sentiment-related objectives in training (Xu et al., 2019; Yin et al., 2020; Zhou et al., 2020). They successively achieved state-of-the-art performance in sentiment analysis tasks among language representation models.

### 2.4 Learning Objectives for ERC Models

ERC is often considered a single-label sequential classification problem. Using softmax cross-entropy loss has been the norm in the training of deep learning ERC models for categorical emotions (Poria et al., 2017; Zhong et al., 2019; Ghosal et al., 2020; Kim and Vossen, 2021) or quantised emotion dimensions (Cerisara et al., 2018; Wang et al., 2020). However, this simplistic cross-entropy loss

ignores the inter-class relations and output probabilities on incorrect classes.

Chen et al. (2019) proposed to suppress the output probabilities of incorrect classes equally while minimising the standard cross-entropy loss. Hou et al. (2016) proposed squared earth mover’s distance to penalise the misclassifications according to a ground distance matrix that quantifies the dissimilarities between classes for image age estimation and aesthetics estimation.

Although highly suitable for emotions, learning from misclassifications is rarely considered because the distance between emotion classes is hard to quantify. Therefore, we propose to leverage the structured label definition of EmoWOZ to model inter-class similarity.

**Multi-task Learning (MTL)** is a technique for learning tasks in parallel using a shared representation. It aims to improve generalisation by using the information in training signals of related tasks as an inductive bias (Caruana, 1997). In emotion recognition, auxiliary tasks include topic classification (Wang et al., 2020) and personality traits (Li et al., 2021). When co-labels are not available, it is also possible to leverage aspects of emotion for additional labels such as valence-arousal (Kim et al., 2017). In this work, we exploit the valence-elicitor-conduct labels in EmoWOZ for MTL.

## 3 Background

### 3.1 User Emotion Recognition

We formulate the task as recognising one emotion class  $e_t$  from a set of  $n$  discrete emotions  $E = \{e^1, e^2, \dots, e^n\}$  in the user turn  $u_t$ , given a dialogue history  $H_t = [u_t, s_{t-1}, u_{t-1}, \dots, s_1, u_1]$ , where  $s$  denotes system turns and  $u$  denotes user turns. Unlike existing chat-ERC models, which are often built for static analysis on the dialogue as a whole, real-time ERC in ToDs does not consider future utterances in dialogue.

### 3.2 User Satisfaction Prediction

User satisfaction prediction aims to predict one satisfaction level  $c_t$  from a set of  $m$  discrete levels  $C = \{c^1, c^2, \dots, c^m\}$  in the user turn  $u_t$ , given all previous turns  $P_t = [s_{t-1}, H_{t-1}]$ . This task differs from ERC in that the user turn  $u_t$  is not available as a part of model input. Since user satisfaction is highly correlated with the valence aspect in user emotion, this task can also be viewed as

user emotion prediction. This is an important task in building ToD systems and has been used for user simulation and system evaluation (Sun et al., 2021).

#### 4 Emotion Recogniser for Task-oriented Dialogues (ERToD)

In this section, we propose our ERToD framework that adapt chit-chat ERC models to the task-oriented domain, as illustrated in Figure 2.

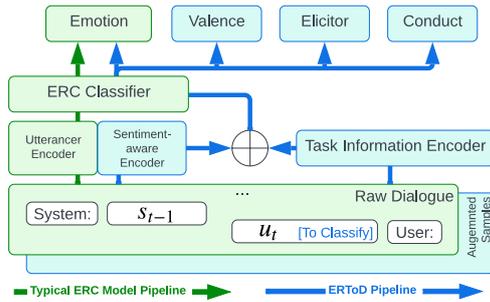


Figure 2: Our proposed ERToD Framework.

##### 4.1 Data Augmentation

Unlike emotions in chit-chat dialogues, resources for emotions in ToDs are very limited. In addition, the data scarcity not only lies in the lack of linguistic diversity but also in the limited domains and actions in which emotions are expressed.

In ToDs, user’s emotional expressions have different degrees of connection to the dialogue task. For example, a user can express dissatisfaction towards the system by pointing out the system’s mistake. In such a case, simply replacing or paraphrasing the user’s utterance based on emotion can potentially break the consistency of the task flow in the context. Such emotions are *context-dependent*.

On the other hand, *context-independent emotions* are expressed without any connection to the user goal, such is the case with abusive utterances. Due to the lack of connection, a simple replacement with a different abusive sentence can fit into the context well without impairing the consistency of task flow in the dialogue.

To obtain augmented samples with meaningful and coherent context, we adopt two different strategies of DA according to the degree of context dependency of emotional expressions.

**Context-independent Emotions** To augment samples for a target emotion  $e$ , we select a user utterance  $u'$  with the equivalent label from other dialogue datasets. We then use it to replace the user utterance  $u_t$  having label  $e$  in the

training data while keeping the original context  $[s_{t-1}, u_{t-1}, \dots, s_1, u_1]$ . The new sample is obtained as  $H'_t = [u', s_{t-1}, u_{t-1}, \dots, s_1, u_1]$ .

**Context-dependent Emotions** We first sample a pool of unlabelled candidate dialogues  $H'_t = [u'_t, s'_{t-1}, u'_{t-1}, \dots, s'_1, u'_1]$  from other ToD datasets. We train a classifier with an uncertainty estimator to identify the emotion label  $e_t$  of the user utterance  $u_t$  and its confidence in each candidate:

$$p(e_t), \text{conf}(e_t) = \text{UncertaintyClassifier}(H'_t) \quad (1)$$

The candidate is selected for emotion  $e_t$  only if  $\text{conf}_t(e)$  is above a confidence threshold  $\theta$ .

##### 4.2 Task Information Encoder

We use a dialogue state tracker (DST) to determine the status of goal completion at each turn. In ToDs, the dialogue state describes the system’s understanding of the the user’s goal up to that point in the dialogue (Young et al., 2010). It encodes dialogue progress in an abstractive manner.

Here as a proof of concept, we use an ontology-dependent DST, which means the concepts that the system can talk about are pre-determined. While we can eliminate the ontology dependency by, for example, using an ontology-independent DST and extracting task features from dialogue state description in natural language, this goes beyond the scope of this work. The DST takes the dialogue history to determine  $\text{SemDS}_t$ , the current dialogue state in semantic form. It is stored as a dictionary that records slots and filled values.  $\text{SemDS}_t$  is then converted into a vector of 0/1’s, indicating whether a particular slot has been filled.

$$V_t = \text{Vectoriser}(\text{SemDS}_t) \quad (2)$$

To account for the change of dialogue state, which depicts how the system performs locally, we concatenate dialogue states of three consecutive turns to obtain a contextual dialogue state vector.

$$\tilde{V}_t = V_t \oplus V_{t-1} \oplus V_{t-2} \quad (3)$$

$V_{t \leq 0}$  are zero vectors, representing the state before the dialogue starts.  $\tilde{V}_t$  is then fed into a trainable fully connected (FC) layer.

$$S_t = \text{FC}(\tilde{V}_t) \quad (4)$$

**Feature Fusion for Emotion Classification** For a chit-chat ERC model with an arbitrary utterance encoder,  $R_t = \text{Encoder}(H_t)$ , i.e.  $R_t$  is the encoded representation of the dialogue history  $H_t$ . The utterance encoder is replaced with a sentiment-aware

encoder in our framework (see Figure 2).

The utterance and the task information encodings are fused via concatenation and fed into the emotion classifier. The output probability of all emotion classes in utterance  $u_t$  is given by:

$$p_t = \text{Softmax}(\text{Classifier}(R_t \oplus S_t)) \quad (5)$$

### 4.3 Learning Objectives

#### 4.3.1 Emotion-Distance Weighted Loss

Emotion classification is a very challenging task due to the subjectivity in the perception of emotion. Since some emotions are more similar to each other than others, it may be advantageous to distinguish marginally wrong recognitions (satisfied vs excited) from extremely wrong ones (satisfied vs dissatisfied). Furthermore, different misclassifications can elicit different user reactions to the dialogue agent. For example, perceiving satisfaction when the user is neutral may or may not annoy the user, but accusing the user of abusive behavior by mistake is a serious offense to the user. Therefore, it is intuitive to penalise misclassifications according to (1) the distance from the label and (2) output probabilities on incorrect labels.

**Defining the Emotion Distance** Since emotion labels in EmoWOZ are defined in three aspects, we can define the distance between emotion labels in terms of their distance on each aspect. A matrix  $D$  is defined where each element  $D(i, j)$  is a vector containing the distance between emotion label  $i$  and  $j$  in each of three aspects (valence, elicitor, and conduct). The matrix  $D$  is symmetric with vector-valued entries.

$$D(i, j) = [d_{val}(i, j), d_{eli}(i, j), d_{con}(i, j)] \quad (6)$$

The final distance is obtained by the sum of the distance in each aspect, followed by an addition of 1 and smoothing with the log operator. The addition of 1 ensures that the log distance is still 0 for identical labels.

$$\tilde{D}(i, j) = \log(\text{sum}(D(i, j)) + 1) \quad (7)$$

#### Considering Misclassification Probabilities

For each sample including the dialogue history  $H_t$ , we look at the softmax output from the model.

$$p_t = \text{Classifier}(H_t) \quad (8)$$

We aim to minimise the probability of each misclassification  $p_t(e = e_i)$  where  $e_i \neq \text{label}_t$ . This is done by maximising  $1 - p_t(e = e_i)$ , the probability of the utterance *not* being wrongly recognised as  $e_i$ . We then calculate the log of this probability so

that in the case of a perfectly correct recognition, the penalty from misclassification will be 0.

$$f(p_t) = \log(1 - p_t) \quad (9)$$

**Obtaining Weights for Misclassifications** We obtain the relevant row in matrix  $D$  that contains the distance between each emotion and the ground-truth label  $j$  of utterance  $u_t$ , followed by a normalisation to obtain a vector  $w_{t,j}$  of normalised emotion-distance weights for all emotions.

$$o_{t,j} = \text{onehot}(\text{label}_t = j) \quad (10)$$

$$\tilde{D}(:, j) = \tilde{D} \times o_{t,j} \quad (11)$$

$$w_{t,j} = \tilde{D}(:, j) / \text{sum}(\tilde{D}(:, j)) \quad (12)$$

**EmoDistLoss** The final loss, which we name *EmoDistLoss*, is calculated from the negative weighted sum of log terms from Equation 9. Since the distance, hence the weight, between identical labels is 0, this calculation does not involve the output probability of the correct label.

$$\text{EmoDistLoss}_t = -w_{t,j} \cdot f(p_t) \quad (13)$$

#### 4.3.2 MTL via Emotional Aspects

In addition to the emotion classification head, we have a classification head for each emotion aspect from the label definition, namely the valence, the elicitor, and the conduct.

The overall classification loss  $L$  is a weighted sum of the loss from softmax outputs of four classification heads  $L_{emo}, L_{val}, L_{eli}, L_{con}$  with a hyperparameter  $\alpha$ .

$$L = \alpha L_{emo} + \frac{1}{3}(1 - \alpha)(L_{val} + L_{eli} + L_{con}) \quad (14)$$

## 5 User Emotion Recognition in ToDs with ERToD

### 5.1 Experimental Set-up

#### 5.1.1 Dataset

We train and test our models on EmoWOZ. It contains user emotion annotations for all dialogues from MultiWOZ (Budzianowski et al., 2018) and additional 1000 human-machine dialogues. It contains 7 emotion groups (see Table 1 and Appendix A for details). Four emotion classes are considerably rare: *fearful*, *apologetic*, *abusive*, and *excited*. DA examples can be found in Appendix B. Our primary aim of DA is to address the poor ERC performance on rare emotions rather than building a balanced dataset. While the later aim can be achieved with the aid of large language models for example, this is out of the scope of our work.

Class Name	Valence	Elicitor	Conduct	Count (%)
<b>Neutral</b>	Neutral	Don't Care	Polite	58,656 (70.1%)
<b>Satisfied</b>	Positive	Operator	Polite	17,532 (21.0%)
<b>Dissatisfied</b>	Negative	Operator	Polite	5,117 (6.1%)
<b>Excited</b>	Positive	Event/Fact	Polite	971 (1.2%)
<b>Apologetic</b>	Negative	User	Polite	840 (1.0%)
<b>Fearful</b>	Negative	Event/Fact	Polite	396 (0.5%)
<b>Abusive</b>	Negative	Operator	Impolite	105 (0.2%)

Table 1: EmoWOZ Emotion definition and distribution.

**Augmenting Abusive Utterances** The user sometimes becomes abusive towards the system. While this correlates with failure to satisfy the user goal, exact abusive expressions uttered by the user are usually independent of the context. Therefore, we apply our DA method for context-independent emotions for *Abusive*. We utilise ConvAbuse, a dataset for nuanced abusive behaviours in chit-chat conversations (Cercas Curry et al., 2021), for more diverse abusive expressions. In ConvAbuse, user utterances are labelled with type, target, strength, and directiveness. We filter for abuses on the system’s intellectuality (labelled as type=intellectual and target=system) to better suit ToD context. We combine each selected utterance with the context of a random abusive utterance in EmoWOZ, resulting in 273 augmented samples.

**Augmenting Fearful, Apologetic, and Excited Utterances** Expressions of these emotions usually contain task information. *Fearful* and *Excited* usually co-occur with a description of the situation that prompts the user to interact with the system. *Apologetic* is frequently associated with a correction of search criteria. There is a strong connection between these emotion expressions and the progression of the task in the dialogue history. Therefore, we apply our DA method for these context-dependent emotions. We look for samples with desired emotions from other ToD datasets using automatic labels. We train a ContextBERT on EmoWOZ (see Section 5.1.2) with a 30% dropout on the BERT output. We train the model with 10 different seeds and run inferences on the training set of existing ToD datasets: Schema-Guided Dialogue (SGD, Rastogi et al. 2019), Taskmaster-1 (TM-1), and Taskmaster-2 (TM-2) (Byrne et al., 2019). In addition, we filter for common domains of EmoWOZ: *Hotels, RideSharing, Travel, Restaurants* in SGD, *RestaurantTable, PizzaOrdering, CoffeeOrdering, UberLyft* in TM-1, and *HotelSearch, Restaurants, FoodOrdering* in TM-2. The classification confidence is measured by votes from 10 models. We use a confidence threshold

of 0.7 and cap the number of augmented samples at 1000 for each emotion, resulting in 268 *fearful*, 872 *apologetic*, and 1000 *excited* samples.

### 5.1.2 Baselines

We implement ERTod to a range of ERC models that have been used to benchmark EmoWOZ, as listed in Table 2. ContextBERT (Feng et al., 2022) and EmoBERTa (Kim and Vossen, 2021) are simple yet robust transformer-based ERC models, and they have similar spirits except that they respectively use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as utterance encoder. They are both built on top of BERT by additionally considering dialogue context and speaker roles in the input. DialogueRNN (Majumder et al., 2019) and COSMIC (Ghosal et al., 2020) are RNN-based models. Following (Feng et al., 2022), we use DialogueRNN with either {GloVe(Pennington et al., 2014)+Convolutional Neural Network} or BERT as the utterance encoder. COSMIC additionally extracts features with a pre-trained commonsense model (Bosselut et al., 2019)<sup>1</sup>. It is important to note that after replacing the original utterance encoder with the sentiment-aware encoder (as described in Section 5.1.3), two variants of DialogueRNN essentially become the same model, and so do EmoBERTa and ContextBERT.

### 5.1.3 Training

In our task information encoder, we use SetSUMBT DST (van Niekerk et al., 2021) from ConvLab-3 toolkit (Zhu et al., 2022). SetSUMBT is a strong DST considering uncertainty with a joint goal accuracy of 52.26% on MultiWOZ 2.1 (Eric et al., 2020). The FC layer in Equation 4 has input/output dimensions of 1083 and 256 respectively and hyperbolic tangent activation (TanH, LeCun et al. 2015). We further replace the utterance encoders of chit-chat ERC models with SentiX, a sentiment-adapted BERT (Zhou et al., 2020).

We use our proposed *EmoDistLoss* for the emotion classification head and cross-entropy loss for MTL heads (valence, elicitor, and conduct). Since the elicitor of *Neutral* emotion is not distinguishable and therefore not explicitly defined in EmoWOZ, we mark the elicitor of *Neutral* samples

<sup>1</sup>COSMIC requires future utterances in recognising the current emotion whereas other models can be configured as either bidirectional or unidirectional. While we use unidirectional set-ups where possible to comply with our task formulation in Section 3.1, we are also interested in how ERTod improves COSMIC for static dialogue analysis in ToDs.

as *don't care*, and their loss in from elicitor classification is ignored.  $\alpha$  in Equation 14 is set to 0.4 based on several rounds of hyperparameter tuning.

To calculate the *EmoDistLoss*, we use 1 as the unit distance and define the distance for each emotional aspect as illustrated in Appendix C. For valence, it is commonly adopted to consider negative and positive as two polarities and neutral in the middle (Socher et al., 2013). Therefore, the distance is 2 between positive and negative, and 1 between non-neutral and neutral. For emotion elicitors, we set the distance between *don't care* to any specific elicitor as 0.5 to penalise a “lazy” classifier that wrongly recognises the emotion as neutral. Doing so also results in a consistent shortest distance of 1 between any pair of specific elicitors.

We follow the default training set-up of each model except for ContextBERT. We reduce the context size of ContextBERT from 512 to 128, resulting in stronger performance and faster training.

### 5.1.4 Evaluation

We report F1 for each emotion. For overall performance, we report both macro F1 and weighted F1. Macro F1 considers each emotion equally and reflects the model’s ability to recognise rare emotions. Weighted F1 is the weighted sum of F1 scores of each label. Weights are determined by the proportion of each emotion in the dataset. We exclude *Neutral* from calculating the averages as it makes up more than 70% of labels.

In addition, we also calculate the average emotion distance (AED) between the recognised emotion and the label to quantify how wrong the model is when it misclassifies. The AED of an emotion  $e$  is calculated from the average of  $\tilde{D}(\text{label}=e, \text{recognised\_emotion})$  of samples whose label is  $e$  (see Equation 7). Lower AED means less severe consequences from mistakes, and is therefore more desirable. All experiments are repeated with 10 different seeds.

## 5.2 ERC Results

Table 2 shows the change in the emotion recognition performance of the selected chat-ERC models after incorporating our ERTToD framework. ERTToD achieves significant improvement in average F1 scores of all models (see Appendix D for examples of model outputs, Appendix E for F1 of individual emotions).

	Base Model		+ ERTToD		Difference	
	MF1	WF1	MF1	WF1	MF1	WF1
BERT	50.1	73.5	61.4	77.3	+11.3	+3.8
DialogueRNN+GloVe	40.1	74.6	56.5	78.5	+16.4	+3.9
DialogueRNN+BERT	52.1	75.5	56.5	78.5	+4.4	+3.0
COSMIC	56.3	77.1	57.4	79.6	+1.1	+2.5
EmoBERTa	57.9	<b>83.0</b>	<b>65.9</b>	<b>83.9</b>	+9.0	+0.9
ContextBERT	<b>59.1</b>	81.9	<b>65.9</b>	<b>83.9</b>	+6.8	+2.0

Table 2: Macro- and weighted-average F1 (MF1, WF1) of ERC models before and after incorporating ERTToD. Best average F1s are marked in **bold**. All differences are significant with  $p < 0.05$ .

	Model	Neu.	Sat.	Dis.	Exc.	Apo.	Fea.	Abu.
F1 Score ( $\uparrow$ )	ContextBERT	93.5	89.1	69.7	45.6	69.6	33.3	47.0
	+ DA	$\uparrow$ <b>94.2</b>	$\uparrow$ 90.5	$\uparrow$ 71.0	45.3	$\uparrow$ 72.1	$\ddagger$ 38.3	$\uparrow$ 67.4
	+ DS	$\uparrow$ <b>94.2</b>	$\uparrow$ 90.5	$\uparrow$ 71.3	45.7	$\uparrow$ 72.7	35.3	$\uparrow$ 69.4
	+ SentiX	$\uparrow$ <b>94.2</b>	$\uparrow$ <b>90.6</b>	$\uparrow$ 72.2	$\ddagger$ 47.1	$\uparrow$ 73.2	$\uparrow$ 39.0	$\uparrow$ 66.1
	+ MTL	$\uparrow$ <b>94.2</b>	$\uparrow$ 90.4	$\uparrow$ <b>72.3</b>	$\ddagger$ 47.2	$\uparrow$ <b>73.4</b>	$\uparrow$ 41.0	$\uparrow$ 67.9
	+ ERTToD	$\uparrow$ 94.1	$\uparrow$ <b>90.6</b>	$\uparrow$ <b>72.3</b>	$\uparrow$ <b>47.6</b>	$\uparrow$ 72.0	$\uparrow$ <b>42.4</b>	$\uparrow$ <b>69.8</b>
	AED Score ( $\downarrow$ )	ContextBERT	0.058	0.094	0.304	0.497	0.269	0.605
+ DA	$\uparrow$ <b>0.049</b>	$\uparrow$ 0.080	0.312	0.493	$\ddagger$ 0.292	0.593	$\uparrow$ 0.339	
+ DS	$\uparrow$ 0.053	$\uparrow$ 0.075	0.296	0.481	0.277	0.582	$\uparrow$ 0.300	
+ SentiX	$\uparrow$ 0.052	$\uparrow$ 0.077	$\ddagger$ 0.286	$\uparrow$ 0.454	0.287	0.596	$\uparrow$ 0.283	
+ MTL	$\uparrow$ 0.054	$\uparrow$ 0.075	$\ddagger$ <b>0.284</b>	$\uparrow$ 0.456	0.277	0.585	$\uparrow$ <b>0.258</b>	
+ ERTToD	0.056	$\uparrow$ <b>0.070</b>	0.296	$\uparrow$ <b>0.435</b>	<b>0.244</b>	<b>0.571</b>	$\uparrow$ 0.277	

Table 3: F1 ( $\uparrow$ ) and AED ( $\downarrow$ ) scores of **Neutral**, **Satisfied**, **Dissatisfied**, **Excited**, **Apologetic**, **Fearful**, and **Abusive**.  $\uparrow$  indicates statistically significant difference with  $p < 0.05$  and  $\ddagger$  indicates  $p < 0.1$  when comparing with ContextBERT. Best scores are marked in **bold**.

## 5.3 Ablation Study on ERTToD

We perform an ablation study on the best performing model, ContextBERT-ERTToD (Table 3). We add each technique in the order of data-related, feature-related, and loss-related approaches. Averaged scores can be found in Appendix F.

**Impact of DA** DA helps improve almost all F1 scores even with a relatively small number of additional samples. There is a small and insignificant drop in the F1 of *Excited*, which is also frequently confused among human annotators. Further work to resolve the ambiguities would be beneficial.

**Impact of Dialogue State (+DS)** Adding dialogue state features further improves most other non-neutral emotions. Although it does not bring advantages for the F1 of *Fearful*, the AED of it continues to improve, showing that the system is making less severe mistakes.

**Impact of SentiX** Initialising BERT with SentiX parameters further improves the recognition of all other non-neutral emotions except for *Abusive*. This suggests that the sentiment information encoded in SentiX is useful for resolving ambiguity. We suspect that, while SentiX is good at distinguishing the valence of emotion, its effect is

limited for user conduct, the hallmark of *Abusive*.

**Impact of MTL** MTL improves F1 for all non-neutral emotions except for *Satisfied*. It also achieves the best AED for *Abusive*. This suggests that MTL heads, especially the conduct classification head, help identify emotions in the simpler valence-elicitor-conduct space. There is a slight drop in the F1 score of *Satisfied*, but it is compensated by the improvement in its AED.

**Impact of *EmoDistLoss* (+ERToD)** The final version of the model achieves the best F1 score in  $\{Satisfied, Dissatisfied, Excited, Fearful, Abusive\}$  and the best AED score in  $\{Satisfied, Excited, Apologetic, Fearful\}$ , leading to best averaged scores (Table F8). This shows penalising misclassifications according to emotion distance, which is only possible thanks to the emotion model, further helps recognise ambiguous emotions.

For the degradation of both scores in *Neutral*, we hypothesise that the model recognises non-neutral emotions more boldly than annotators, who are more cautious about subtle emotional cues.

## 6 Zero-shot User Satisfaction Prediction

### 6.1 Experimental Set-up

#### 6.1.1 Dataset

We evaluate our model with **User Satisfaction Simulation** (USS) dataset where user utterances are annotated with 5-level satisfaction ratings (Sun et al., 2021). Dialogues in USS come from 5 different ToD datasets:

**Jing Dong Dialogue Corpus** (JDDC, Chen et al., 2020) is a multi-turn Chinese dialogue dataset for E-commerce customer service. USS contains 54.5k user satisfaction annotations for 3300 dialogues sampled from JDDC. Since JDDC is in Chinese, we translated it into English with Google Translate API first.

**Schema-guided Dialogues** (SGD, Rastogi et al., 2020) is a multi-domain, task-oriented conversations between a human and a virtual assistant. These conversations involve interactions with services and APIs spanning 20 domains, such as banks, events, media, calendar, travel, and weather. USS contains 13.8k user satisfaction annotations for 1000 dialogues sampled from SGD. Although we use SGD for DA, our DA samples do not overlap with SGD dialogues in USS.

**Recommendation Dialogue** (ReDial, Li et al., 2018) is an annotated dataset of dialogues, where users recommend movies to each other. USS contains 11.8k user satisfaction annotations for 1000 dialogues sampled from ReDial.

**Coached Conversational Preference Elicitation** (CCPE, Radlinski et al., 2019) is a dialogue dataset where the “assistant” is tasked with eliciting the “user” preferences about movies collected in the Wizard-of-Oz framework. USS contains 6.8k user satisfaction annotations for 500 dialogues sampled from CCPE.

**MultiWOZ** (Budzianowski et al., 2018) is a multi-domain task-oriented dialogue dataset collected in the Wizard-of-Oz framework spanning 7 domains such as restaurant, hotel, and attraction. USS contains 12.5k user satisfaction annotations for 1000 dialogues sampled from MultiWOZ. Since we trained our ERC model on EmoWOZ, which was based on MultiWOZ, we excluded it in our evaluation.

#### 6.1.2 Baselines

We compare our zero-shot results with supervised models of Sun et al. (2021) and Kim and Lipani (2022). HiGRU (Yang et al., 2016) and BERT (Devlin et al., 2019) were the best two models trained by Sun et al. (2021) to benchmark USS dataset when it was first released. SatAct and SatActUtt are T5-based models (Raffel et al., 2020). SatAct is trained to predict user satisfaction and user action in a MTL set-up, whereas SatActUtt additionally incorporates user utterance generation. For satisfaction prediction, these models were set up to predict a 5-level rating during training.

These baseline models were trained on each one of the five ToD subsets in USS with a 10-fold cross-validation. Although non-3 ratings were up-sampled by 10 times in their training, the training data size is still smaller than that of ContextBERT-ERToD (68.9k emotion annotations, EmoWOZ and DA samples altogether).

#### 6.1.3 Zero-shot Inference

We experimented with ContextBERT-ERToD, the best resulting model from ERC training. After training the model for ERC, we fixed its parameters and ran inference with USS dataset for zero-shot user satisfaction prediction. To adapt to user satisfaction prediction set-up, we excluded information about the user turn at  $t$  from the model

input as well as the dialogue state. Specifically, for utterance encoding, we excluded  $u_t$  from the dialogue history to have  $H_t = [s_{t-1}, u_{t-1}, \dots, s_1, u_1]$ . For task information encoding, we shifted the context window in Equation 3 by one and have  $\tilde{V}_t = V_{t-1} \oplus V_{t-2} \oplus V_{t-3}$  as the new contextual dialogue state vector.

#### 6.1.4 Evaluation

In the works of baseline models, satisfaction ratings {1,2} were considered the negative class and {3,4,5} as the positive. To map the emotion prediction from our ERC model to binary satisfaction ratings, it is intuitive to leverage the valence aspect of emotions. Emotion classes with a negative valence were considered *Not Satisfied* and those with a positive valence as *Satisfied*. The emotion *Apologetic* is an exception among emotions with a negative valence. Since its elicitor is the user him/herself, it should not be considered as a sign of user dissatisfaction. Regarding the emotion class *Neutral*, we mapped it to *Satisfied* because the original evaluation set-up of baseline models considered the medium satisfaction rating, 3, as the positive class.

Overall, we considered {*Neutral, Apologetic, Excited, Satisfied*} as the positive class and {*Fearful, Dissatisfied, Abusive*} as negative.

## 6.2 Results

	JDDC	SGD	ReDial	CCPE
HiGRU (Sun et al., 2021)	17.1	8.6	8.3	27.4
BERT (Sun et al., 2021)	18.5	4.8	12.5	24.5
SatAct (Kim and Lipani, 2022)	-	71.3	-	16.5
SatActUtt (Kim and Lipani, 2022)	-	<b>84.7</b>	-	73.4
ContextBERT-ERToD (0-shot)	<b>50.8</b>	78.8	<b>78.1</b>	<b>77.6</b>

Table 4: Binary F1 scores on different USS subsets. Best scores are marked in **bold**.

Following existing work, we first report binary F1 for direct comparison. In Table 4, ContextBERT-ERToD performs comparably with SatActUtt and significantly outperforms other models. This shows that our ERToD framework in combination with the ERC model generalises well to user satisfaction prediction.

## 7 Conclusion

In this work, we propose ERToD, a framework to address three critical steps in learning and effectively adapt chit-chat ERC models to recognise emotions in ToDs. We propose two strategies of

DA for different emotions to improve ERC performance in ToDs on rare emotions. We further leverage dialogue state and sentiment-aware embeddings for a richer feature representation. In addition, we apply MTL and devise a novel loss function, *EmoDistLoss*, which take the similarities between emotions into account. Our framework significantly improves existing chit-chat ERC models’ performance in recognising user emotions in ToDs. By further applying our best resulting model to perform the task of user satisfaction prediction, we show that our method generalises well on other similar valence-related classification tasks in ToDs.

As more sophisticated and powerful dialogue systems such as ChatGPT arise, there is an urge to recognise, understand and handle the emotion of the user, especially in the age where online abuse is omnipresent. The long-term aim of this work is to obtain valuable insight for downstream ToD modelling tasks. This allows further investigation of emotion regulation strategies on the system side to improve task performance and user satisfaction, and to prevent undesirable user behaviours.

## 8 Acknowledgements

S. Feng, N. Lubis, M. Heck, and C. van Niekerk are supported by funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research, while C. Geishauser, H-C. Lin, B. Ruppik, and R. Vukovic are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018804636). Computing resources were provided by Google Cloud.

## References

- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. *COMET: Commonsense transformers for automatic knowledge graph construction*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. *MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

- Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. **Taskmaster-1: Toward a realistic and diverse dialog dataset**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.
- Rich Caruana. 1997. **Multitask learning**. *Machine Learning*, 28(1):41–75.
- Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. **ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. **Multi-task dialog act and sentiment recognition on mastodon**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. 2019. **Complement objective training**. In *International Conference on Learning Representations*.
- Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. **The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 459–466, Marseille, France. European Language Resources Association.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *International Conference on Machine Learning*.
- Niko Colneric and Janez Demšar. 2020. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11:433–446.
- L. Devillers, L. Lamel, and I. Vasilescu. 2003. **Emotion detection in task-oriented spoken dialogues**. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 3, pages III–549.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. **MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.
- Shutong Feng, Nurul Lubis, Christian Geischauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. **EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.
- Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. **COSMIC: COMmonSense knowledge for eMotion identification in conversations**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.
- Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2021. **Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management**. *Transactions of the Association for Computational Linguistics*, 9:36–52.
- Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geischauser, Hsien-Chin Lin, and Milica Gašić. 2022. **Robust Dialogue State Tracking with Weak Supervision and Sparse Data**. *Transactions of the Association for Computational Linguistics*, 10:1175–1192.
- Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. **Squared earth mover’s distance-based loss for training deep neural networks**. *ArXiv*, abs/1611.05916.
- Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. **Sequence-to-sequence data augmentation for dialogue language understanding**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New

- Mexico, USA. Association for Computational Linguistics.
- Zhiqiang Hu, Roy Kaa-Wei Lee, and Nancy F. Chen. 2022. [Are current task-oriented dialogue systems able to satisfy impolite users?](#) *ArXiv*, abs/2210.12942.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. [TinyBERT: Distilling BERT for natural language understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.
- Nam Kyun Kim, Jiwon Lee, Hun Kyu Ha, Geon Woo Lee, Jung Hyuk Lee, and Hong Kook Kim. 2017. [Speech emotion recognition based on multi-task learning using a convolutional neural network](#). In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 704–707.
- Taewoon Kim and Piek Vossen. 2021. [EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa](#). *ArXiv*, abs/2108.12009.
- To Eun Kim and Aldo Lipani. 2022. [A multi-task based neural model to simulate users in goal oriented dialogue systems](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, page 2115–2119, New York, NY, USA. Association for Computing Machinery.
- Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. 2019. [A closer look at feature space data augmentation for few-shot intent classification](#).
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.
- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. [Towards deep conversational recommendations](#). In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.
- Yang Li, Amirmohammad Kazameini, Yash Mehta, and Erik Cambria. 2021. [Multitask learning for emotion and personality detection](#). *ArXiv*, abs/2101.02346.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2015. [Construction and analysis of social-affective interaction corpus in english and indonesian](#). In *2015 International Conference Oriental COCODA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCODAS/CASLRE)*, pages 202–206. IEEE.
- Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. [DialogueRNN: An attentive RNN for emotion detection in conversations](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.
- Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. [The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent](#). *IEEE Transactions on Affective Computing*, 3(1):5–17.
- Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023. [Genpads: Reinforcing politeness in an end-to-end dialogue system](#). *PLOS ONE*, 18(1):1–20.
- Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. [A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5).
- Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. [Context-dependent sentiment analysis in user-generated videos](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.
- Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. [Coached conversational preference elicitation: A case study in understanding](#)

- movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360, Stockholm, Sweden. Association for Computational Linguistics.
- Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. 2017. Data distillation: Towards omni-supervised learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI Conference on Artificial Intelligence*.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel. *Proceedings of the VLDB Endowment*, 11(3):269–282.
- Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PLOS ONE*, 15(7):1–28.
- Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let’s go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3369–3373, Istanbul, Turkey. European Language Resources Association (ELRA).
- Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia. Association for Computational Linguistics.
- Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837, Marseille, France. European Language Resources Association.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, page 2499–2506, New York, NY, USA. Association for Computing Machinery.
- Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.
- Carel van Niekerk, Andrey Malinin, Christian Geisshauer, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. Uncertainty measures in neural belief tracking and the effects on dialogue policy performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184.
- Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020a. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA. Curran Associates Inc.

- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. [BERT post-training for review reading comprehension and aspect-based sentiment analysis](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yifei Xu, Jingqiao Zhang, Ru He, Liangzhu Ge, Chao Yang, Cheng Yang, and Ying Nian Wu. 2022. [Sas: Self-augmentation strategy for language model pre-training](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11586–11594.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. [Hierarchical attention networks for document classification](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.
- Da Yin, Tao Meng, and Kai-Wei Chang. 2020. [SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online. Association for Computational Linguistics.
- Steve Young. 2002. Talking to machines (statistically speaking). In *Seventh International Conference on Spoken Language Processing*.
- Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. [The hidden information state model: A practical framework for POMDP-based spoken dialogue management](#). *Computer Speech & Language*, 24(2):150–174.
- Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. [Refining word embeddings for sentiment analysis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.
- Sayyed Zahiri and Jinho D. Choi. 2018. [Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks](#). In *Proceedings of the AAAI Workshop on Affective Content Analysis, AFFCON’18*, pages 44–51, New Orleans, LA.
- Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. [Knowledge-enriched transformer for emotion detection in textual conversations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.
- Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. [Emotional chatting machine: Emotional conversation generation with internal and external memory](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. [SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Qi Zhu, Christian Geisshauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. [Convlab-3: A flexible dialogue system toolkit based on a unified data format](#).

## A Emotion Definitions in EmoWOZ

Elicitor	Valence	Conduct	OCC Emotion Tokens	EmoWOZ Emotion	Implication of User
Operator	Positive	Polite	Admiration, gratitude, love	<i>Satisfied</i> , liking, appreciative	Satisfied with the operator because the goal is fulfilled.
		Impolite		Not applicable to EmoWOZ	
Operator	Negative	Polite	Reproach, anger, hate	<i>Dissatisfied</i> , disliking	Dissatisfied with the operator's suggestion or mistake.
		Impolite		<i>Abusive</i>	Insulting the operator when the goal is not fulfilled.
User	Positive	Polite	Pride, gratification	Not applicable to EmoWOZ	
		Impolite			
User	Negative	Polite	Shame, remorse, hate	<i>Apologetic</i>	Apologising for causing confusion to the operator.
		Impolite		Not modelled in EmoWOZ	Insulting the operator for no reason.
Events, facts	Positive	Polite	Happy-for, gloating, love, satisfaction, relief, joy	<i>Excited</i> , happy, anticipating	Looking forward to a good event (e.g. birthday party).
		Impolite		Not applicable to EmoWOZ	
Events, facts	Negative	Polite	Distress, resentment, hate, fears-confirmed, pity, disappointment	<i>Fearful</i> , sad, disappointed	Encountered a bad event (e.g. robbery and option not available).
		Impolite		Not applicable to EmoWOZ	
-	Neutral	Polite	-	<i>Neutral</i>	Describing situations and needs.
		Impolite		Not modelled in EmoWOZ	No emotion but rude (e.g. using imperative sentences).

Table A1: EmoWOZ labels and similar emotions tokens from the OCC emotion model. For simplicity, emotion words in blue are used to represent each emotion category.

## B Examples of Augmented Samples

### B.1 Augmentation with Automatic Label

Source: Taskmaster-1 Dialogue ID: dlg-02edb443-9d6f-4553-af6e-f69778eb0fc5
...
S: Any other restaurant you were thinking about?
U: Yes, what about Char's at Tracy Mansion?
S: You mean the one on 1829 N Front St, Harrisburg, PA 17102-2213
U: <b>Yes, I've heard great things about that</b> [DA Candidate]
ContextBERT Ensemble Prediction: (Emotion = Excited, Confidence = 100%)

Figure B.1: DA sample for emotion *Excited*.

Source: SGD Dialogue ID: 93_00124
...
U: I plan to check in on the 7th of this month.
S: Okay. You want to reserve 1 room at 1 Hotel Brooklyn Bridge in New York, for a 1 day stay on March 7th?
U: <b>Sorry, I misremembered the date. I actually want to check into one room on Saturday this week.</b> [DA Candidate]
ContextBERT Ensemble Prediction: (Emotion = Apologetic, Confidence = 100%)

Figure B.2: DA sample for emotion *Apologetic*.

Source: SGD Dialogue ID: 1_00028
U: <b>Hey! Today feeling so tired... Wish to have some food outside in a good restaurant... Can you find me one?</b> [DA Candidate] ...
ContextBERT Ensemble Prediction: (Emotion = Fearful, Confidence = 90%), (Emotion = Excited, Confidence = 10%)

Figure B.3: DA sample for emotion *Fearful*. Please note that although this class is called “fearful” for simplicity, user’s negative emotion due to any undesirable events that is out of the control of the operator also belongs to this category in EmoWOZ according to Table A1.

## B.2 Augmentation with Existing Dataset and Utterance Replacement

Source: ConvAbuse
DA Candidates:
1. <b>Are you an idiot</b> 2. <b>You are stupid privitive word-repeater</b> 3. <b>Why do you not answer my question</b> 4. <b>Foolish</b> ..
EmoWOZ sample for <i>Abusive</i> Dialogue ID: DMAGE2157.json
... U: need a good resturant near macy's in nyc S: You are more than welcome! U: can you help S: The address is Parkside, Cambridge . The phone is 01223358966 . The postcode is none . Thank you for using our services . U: <b>your horrible and not helpful at all</b> [Label = Abusive]

Figure B.4: DA sample for emotion *Abusive*. Candidate DA samples from ConvAbuse can be used to replace the abusive user utterance in EmoWOZ without causing any context inconsistency.

## C Emotional Aspect Distance Definition

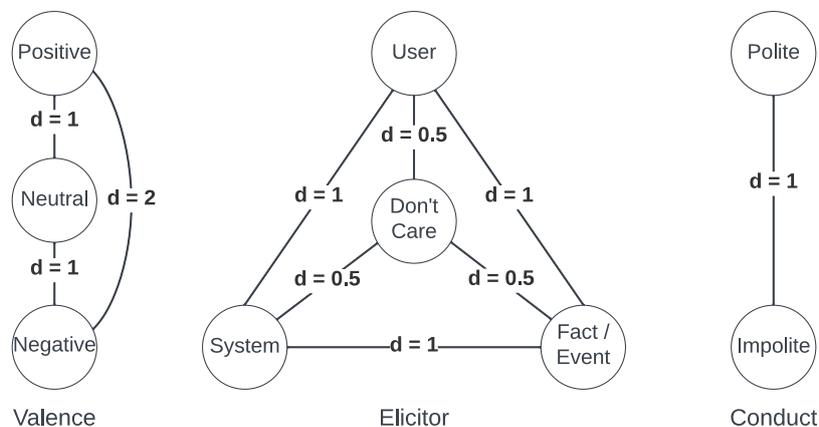


Figure C.1: Distance definition for Equation 6.

## D Examples of Model Recognitions

Dialogue ID: DMAGE3777.json						
...						
U: Can you help me?						
S: The phone is 01223358966. The postcode is none.						
U: Ok						
S: The address is Parkside, Cambridge.						
U: <b>Nice</b> [To classify, label = Dissatisfied]						
	BERT	DialRNN-G	DialRNN-B	COSMIC	EmoBERTa	ContextBERT
Base	Satisfied	Dissatisfied	Satisfied	Satisfied	Satisfied	Neutral
+ ERToD	Neutral	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied

Figure D.1: Model Recognitions on dialogue DMAGE3777 in EmoWOZ.

Dialogue ID: PMUL2437.json						
...						
S: There are 21 restaurants available in the centre of town. How about a specific type of cuisine?						
U: I need to know the food type and postcode and it should also have multiple sports						
S: I am sorry I do not understand what you just said. Please repeat in a way that makes sense.						
U: <b>Get me the food type and the post code</b> [To classify, label=Dissatisfied]						
	BERT	DialRNN-G	DialRNN-B	COSMIC	EmoBERTa	ContextBERT
Base	Neutral	Dissatisfied	Neutral	Neutral	Dissatisfied	Neutral
+ ERToD	Neutral	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied

Figure D.2: Model Recognitions on dialogue PMUL2437 in EmoWOZ

## E Detailed ERC Performance on Each Emotion

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	89.8	88.8	35.1	42.9	70.4	36.2	27.5
DialogueRNN+GloVe	83.5	86.4	51.4	32.7	57.7	12.7	0.0
DialogueRNN+BERT	86.9	87.6	47.5	39.4	<b>71.5</b>	41.3	25.6
COSMIC	89.8	88.4	50.7	44.4	70.9	<b>52.0</b>	31.6
EmoBERTa	<b>94.0</b>	<b>90.3</b>	<b>71.0</b>	44.9	70.6	31.3	39.3
ContextBERT	93.5	89.1	69.7	<b>45.6</b>	69.6	33.3	<b>47.0</b>

Table E2: F1 scores of selected chat-ERC models BEFORE incorporating ERToD framework. The best score for each emotion is marked in **bold**.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	90.3	89.3	88.4	89.2	38.9	38.6	<b>47.7</b>	39.1	69.7	71.5	47.7	30.0	42.1	22.4	55.7	48.5	74.5	74.5
DialRNN-GloVe	<b>97.6</b>	73.0	78.5	<b>95.9</b>	36.5	<b>87.6</b>	22.2	<b>65.7</b>	44.7	<b>82.5</b>	11.2	18.9	0	0	32.2	<b>58.4</b>	65.0	<b>91.4</b>
DialRNN-BERT	94.0	80.7	84.7	90.7	34.8	75.3	36.5	42.9	68.3	75.0	46.7	37.5	28.6	23.5	49.9	57.5	70.4	84.2
COSMIC	93.1	86.8	86.2	90.7	42.3	64.4	43.7	45.3	<b>71.9</b>	70.1	<b>65.0</b>	<b>43.3</b>	<b>77.3</b>	20.0	<b>64.4</b>	55.6	74.0	81.7
EmoBERTa	94.2	<b>94.0</b>	<b>88.7</b>	92.2	<b>74.6</b>	69.5	45.6	42.6	73.0	70.3	37.9	27.2	54.0	24.7	62.3	54.4	<b>82.9</b>	83.8
ContextBERT	93.4	93.7	88.5	89.8	72.6	67.2	46.4	45.4	68.3	71.6	37.9	30.0	64.5	<b>37.6</b>	63.0	57.0	82.3	81.8

Table E3: Precision and Recall scores of selected chit-chat ERC models BEFORE incorporating ERToD framework. We report scores of each emotion: **Neutral**, **Satisfied**, **Dissatisfied**, **Excited**, **Apologetic**, **Fearful**, **Abusive**, as well as Macro- and Weighted Averaged scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	92.4	90.4	43.7	<b>49.7</b>	75.4	39.5	<b>69.7</b>
DialogueRNN+GloVe	92.6	90.1	51.4	43.9	<b>77.6</b>	42.4	33.8
DialogueRNN+BERT	92.6	90.1	51.4	43.9	<b>77.6</b>	42.4	33.8
COSMIC	91.1	89.5	58.1	45.6	73.3	36.3	41.6
EmoBERTa	<b>94.0</b>	<b>90.5</b>	<b>72.3</b>	47.9	71.9	<b>43.4</b>	<b>69.7</b>
ContextBERT	<b>94.0</b>	<b>90.5</b>	<b>72.3</b>	47.9	71.9	<b>43.4</b>	<b>69.7</b>

Table E4: F1 scores of selected chit-chat ERC models AFTER incorporating ERToD framework. The best score for each emotion is marked in **bold**.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	91.0	93.8	88.9	92.0	57.5	35.5	<b>51.2</b>	48.9	<b>81.6</b>	70.3	48.1	33.9	<b>74.8</b>	65.9	67.0	57.7	79.8	76.3
DialRNN-GloVe	91.3	<b>94.0</b>	<b>89.7</b>	90.5	60.9	41.5	44.4	45.6	76.5	77.3	42.6	<b>38.3</b>	54.3	30.0	61.4	53.9	80.6	76.5
DialRNN-BERT	91.3	<b>94.0</b>	<b>89.7</b>	90.5	60.9	41.5	44.4	45.6	76.5	77.3	42.6	<b>38.3</b>	54.3	30.0	61.4	53.9	80.6	76.5
COSMIC	<b>94.4</b>	88.3	86.9	92.3	51.6	<b>68.9</b>	38.7	<b>57.4</b>	68.2	<b>79.3</b>	36.2	<b>38.3</b>	44.7	38.8	54.4	62.5	75.9	<b>84.6</b>
EmoBERTa	94.3	93.9	88.9	<b>92.4</b>	<b>75.6</b>	68.0	45.7	50.7	70.8	74.4	<b>54.6</b>	35.6	72.4	<b>68.2</b>	<b>68.0</b>	<b>64.9</b>	<b>83.5</b>	84.3
ContextBERT	94.3	93.9	88.9	<b>92.4</b>	<b>75.6</b>	68.0	45.7	50.7	70.8	74.4	<b>54.6</b>	35.6	72.4	<b>68.2</b>	<b>68.0</b>	<b>64.9</b>	<b>83.5</b>	84.3

Table E5: Precision and Recall scores of selected chit-chat ERC models AFTER incorporating ERToD framework. We report scores of each emotion: **Neutral**, **Satisfied**, **Dissatisfied**, **Excited**, **Apologetic**, **Fearful**, **Abusive**, as well as Macro- and Weighted Averaged scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	+2.6	+1.6	+8.6	+6.8	+5.0	+3.3	+42.2
DialogueRNN+GloVe	+9.1	+3.7	+0.0	+11.2	+19.9	+29.7	+33.8
DialogueRNN+BERT	+5.7	+2.5	+3.9	+4.5	+6.1	+1.1	+8.2
COSMIC	+1.3	+1.1	+7.4	+1.2	+2.4	<b>-15.7</b>	+10.0
EmoBERTa	0.0	+0.2	+1.3	+3.0	+1.3	+12.1	+30.4
ContextBERT	+0.5	+1.4	+2.6	+2.3	+2.3	+10.1	+22.7

Table E6: Change of F1 scores of selected chit-chat ERC models after incorporating ERToD framework. The only degradation in performance is marked in **bold**.

In terms of F1 scores, ERToD results in improvement in all emotions except for *fearful* in COSMIC (Table E6). We further investigate this exception. While most of fearful utterances are located at the beginning

of the dialogue in the training and development set in EmoWOZ, the position of such utterances are more evenly distributed in the test set as well as the augmented samples. Upon toggling the development set and the test set for evaluation, we observe that the F1 of fearful by COSMIC drops significantly (52.0%  $\rightarrow$  28.8%) while that of COSMIC-ERToD remains roughly unchanged (35.5%  $\rightarrow$  37.6%). The trend in all other results remains unchanged.

The drastically different performance of COSMIC on the development and the test set suggests that COSMIC develops a positional bias from the training set of EmoWOZ. At the same time, COSMIC-ERToD performs similarly on both non-training sets, likely relying more on textual and task information. The limited performance of COSMIC-ERToD is likely due to the extra false-positives at the later stage of dialogues.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	+0.7	+4.5	+0.5	+2.8	+18.6	-3.1	+3.5	+9.8	+11.9	-1.2	+0.4	+3.9	+32.7	+43.5	+11.3	+9.2	+5.3	+1.8
DialRNN-GloVe	-6.3	+21.0	+11.2	-5.4	+24.4	-46.1	+22.2	-20.1	+31.8	-5.2	+31.4	+19.4	+54.3	+30.0	+29.2	-4.5	+15.6	-14.9
DialRNN-BERT	-2.7	+13.3	+5.0	-0.2	+26.1	-33.8	+7.9	+2.7	+8.2	+2.3	-4.1	+0.8	+25.7	+6.5	+11.5	-3.6	+10.2	-7.7
COSMIC	+1.3	+1.5	+0.7	+1.6	+9.3	+4.5	-5.0	+12.1	-3.7	+9.2	-28.8	-5.0	-32.6	+18.8	-10.0	+6.9	+1.9	+2.9
EmoBERTa	+0.1	-0.1	+0.2	+0.2	+1.0	-1.5	+0.1	+8.1	-2.2	+4.1	+16.7	+8.4	+18.4	+43.5	+5.7	+10.5	+0.6	+0.5
ContextBERT	+0.9	+0.2	+0.4	+2.6	+3.0	+0.8	-0.7	+5.3	+2.5	+2.8	+16.7	+5.6	+7.9	+30.6	+5.0	+7.9	+1.2	+2.5

Table E7: The difference in **Precision** and **Recall** scores of selected chit-chat ERC models before and after incorporating ERToD framework. We report scores of each emotion: **Neutral**, **Satisfied**, **Dissatisfied**, **Excited**, **Apologetic**, **Fearful**, **Abusive**, as well as **Macro-** and **Weighted Averaged** scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

## F Averaged Scores for the Ablation Study

	Model	Macro Avg	Weighted Avg
F1 Score ( $\uparrow$ )	ContextBERT	59.1	81.9
	+ DA	$\dagger$ 64.1	$\dagger$ 83.4
	+ DS	$\dagger$ 64.1	$\dagger$ 83.5
	+ SentiX	$\dagger$ 64.8	$\dagger$ 83.7
	+ MTL	$\dagger$ 65.3	$\dagger$ 83.7
	+ ERToD	$\dagger$ <b>65.7</b>	$\dagger$ <b>83.9</b>
AED Score ( $\downarrow$ )	ContextBERT	0.387	0.168
	+ DA	$\dagger$ 0.351	$\dagger$ 0.159
	+ DS	$\dagger$ 0.335	$\dagger$ 0.151
	+ SentiX	$\dagger$ 0.331	$\dagger$ 0.149
	+ MTL	$\dagger$ 0.322	$\dagger$ 0.147
	+ ERToD	$\dagger$ <b>0.316</b>	$\dagger$ <b>0.145</b>

Table F8: Ablation Study of ERToD.  $\dagger$  indicates statistically significant difference with  $p < 0.05$  when comparing with ContextBERT. The best score in each category is in **bold**. For each of the additional methods: DA = Data Augmentation, DS = Dialogue State Features, SentiX = Sentiment-aware Text Embedding, MTL = Multi-task Learning. Neutral is excluded when calculating the averaged scores.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
ContextBERT	93.4	93.9	88.5	<b>92.4</b>	72.6	68.0	46.4	50.7	68.3	74.4	37.9	<b>35.6</b>	64.5	68.2	63.0	64.9	82.3	84.3
+ DA	93.9	94.4	89.4	91.6	75.6	67.2	47.2	44.6	75.0	70.0	53.1	30.6	70.1	65.9	<b>68.4</b>	61.6	84.0	83.1
+ DS	93.8	<b>94.6</b>	<b>90.1</b>	90.9	74.5	68.4	<b>47.9</b>	44.6	75.8	69.0	50.7	27.8	69.9	69.4	68.1	61.7	<b>84.2</b>	82.9
+ SentiX	94.1	94.3	89.5	91.7	76.0	69.1	47.5	49.3	<b>76.7</b>	70.3	50.9	32.2	66.0	66.5	67.8	63.2	84.1	83.9
+ MTL	94.2	94.0	88.9	91.5	<b>76.4</b>	<b>70.6</b>	45.7	<b>49.8</b>	76.6	<b>71.6</b>	51.2	35.0	67.0	<b>72.4</b>	67.6	<b>65.1</b>	83.8	<b>84.2</b>
+ ERToD	<b>94.3</b>	94.1	88.9	91.9	75.6	69.3	45.7	48.8	70.8	70.8	<b>54.6</b>	34.4	<b>72.4</b>	70.0	68.0	64.2	83.5	84.1

Table F9: Ablation study on **Precision** and **Recall** scores of ERToD. We report scores of each emotion: **Neutral**, **Satisfied**, **Dissatisfied**, **Excited**, **Apologetic**, **Fearful**, **Abusive**, as well as **Macro-** and **Weighted Averaged** scores. The best score for each emotion is marked in **bold**. For each of the additional methods: DA = Data Augmentation, DS = DialogueState Features, SentiX = Sentiment-aware Text Embedding, MTL = Multi-task Learning.. Neutral is excluded when calculating averaged scores.

## Chapter 7

# Affect Recognition in Conversations Using Large Language Models

This chapter summarises our work on investigating the ability of large language models to recognise various types of affects under different dialogue settings, and gives a verbatim copy of our manuscript (Feng et al., 2024b):

Shutong Feng et al. (Sept. 2024b). “Affect Recognition in Conversations Using Large Language Models”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, pp. 259–273. DOI: 10.18653/v1/2024.sigdial-1.23. URL: <https://aclanthology.org/2024.sigdial-1.23/>

## 7.1 Summary

In Chapters 5 and 6, we introduce a dedicated corpus for studying user emotions in task-oriented dialogues and propose methods for optimising emotion recognition in such contexts. These efforts involved substantial manual annotation, data curation, and domain-related feature engineering.

With the emergence of large language models (LLMs) that demonstrate strong instruction-following capabilities, we became interested in their potential for affect recognition with minimal task-specific supervision. This chapter explores how effectively LLMs can recognise human affect compared to state-of-the-art, fine-tuned models.

Specifically, we evaluated LLMs on three distinct affect recognition tasks: emotion recognition in open-domain conversations (IEMOCAP (Busso et al., 2008)), emotion recognition in task-oriented dialogues (EmoWOZ (Feng et al., 2022)), and detection of depressive mental states in clinical interviews (DAIC-WOZ (Gratch et al., 2014)). We benchmarked a variety of open-source and proprietary LLMs under different inference setups, including zero-shot prompting, few-shot in-context learning, and supervised fine-tuning on varying portions of each dataset. Additionally, we simulated a realistic deployment scenario by introducing automatic speech recognition (ASR) errors, assessing the robustness of LLMs to noisy input.

Our findings reveal the limited performance of LLMs compared with small, lightweight models with task-specific optimisation. Yet, we also show the promising potential of LLMs in affect recognition, particularly in low-resource or noisy settings.

## 7.2 Personal Contributions

The implementation, technical results, and writing are my own work. Guangzhi Sun assisted in coding. Wen Wu provided the state-of-the-art result on the DAIC-WOZ test set. Milica Gasić supervised the project. Co-authors assisted in writing and proofreading.

# Affect Recognition in Conversations Using Large Language Models

Shutong Feng<sup>1</sup>, Guangzhi Sun<sup>2</sup>, Nurul Lubis<sup>1</sup>, Wen Wu<sup>2</sup>, Chao Zhang<sup>3</sup>, Milica Gašić<sup>1</sup>

<sup>1</sup>Heinrich Heine University Düsseldorf, Germany

<sup>2</sup>University of Cambridge, UK

<sup>3</sup>Tsinghua University, China

{fengs, lubis, gasic}@hhu.de<sup>1</sup>, {gs534, ww368}@cam.ac.uk<sup>2</sup>, cz277@tsinghua.edu.cn<sup>3</sup>

## Abstract

Affect recognition, encompassing emotions, moods, and feelings, plays a pivotal role in human communication. In the realm of conversational artificial intelligence, the ability to discern and respond to human affective cues is a critical factor for creating engaging and empathetic interactions. This study investigates the capacity of large language models (LLMs) to recognise human affect in conversations, with a focus on both open-domain chit-chat dialogues and task-oriented dialogues. Leveraging three diverse datasets, namely IEMOCAP (Busso et al., 2008), EmoWOZ (Feng et al., 2022), and DAIC-WOZ (Gratch et al., 2014), covering a spectrum of dialogues from casual conversations to clinical interviews, we evaluate and compare LLMs’ performance in affect recognition. Our investigation explores the zero-shot and few-shot capabilities of LLMs through in-context learning as well as their model capacities through task-specific fine-tuning. Additionally, this study takes into account the potential impact of automatic speech recognition errors on LLM predictions. With this work, we aim to shed light on the extent to which LLMs can replicate human-like affect recognition capabilities in conversations.

## 1 Introduction

Affect refers to the broad range of subjective experiences related to emotions, moods, and feelings (Russell, 1980). It encompasses the various ways individuals perceive, experience, and express their emotional states and is an essential aspect of human experience and communication (Gross, 2002).

The ability to recognise human affect is an important ability of conversational artificial intelligence (AI, Mayer et al. 1999). It empowers the dialogue agent to go beyond mere information exchange and engage users on an emotional level. By leveraging affect recognition techniques, they can discern the emotional nuances in user inputs, including

sentiment, mood, and subtle cues like sarcasm or frustration (Picard, 1997). This capability allows the system to respond with greater sensitivity, empathy, and relevance, leading to more meaningful and satisfying interactions (Zeng et al., 2009).

Large language models (LLMs) have demonstrated promising performance in many tasks (Beeching et al., 2023). They have also shown promising capability in adapting to new tasks via prompting (Heck et al., 2023; Sun et al., 2023), in-context learning (ICL, Zhao et al. 2023), as well as task-specific fine-tuning (Taori et al., 2023). With the advancement in LLMs, it is possible to use LLMs as the backend of dialogue systems (OpenAI, 2022, 2023; Touvron et al., 2023b). This brings up the question: can LLMs recognise human affects in conversations in a similar capacity as human beings?

In the context of conversational AI, dialogues can be broadly categorised into two classes: 1) chit-chat or open-domain dialogues where users interact with the system for entertainment and engagement, and 2) task-oriented dialogues (ToDs) where users converse with the system for specific goals (Jurafsky and Martin, 2009). Under ToDs, depending on the type of user goals, dialogues can be further grouped as information-retrieval, medical consultations, education, and many more.

Regarding the affective information in conversations, we are particularly interested in the following: (1) categorical emotion classes from generic emotion models such as “basic emotions” proposed by Ekman and Friesen (1971), (2) custom categorical emotion classes defined for a particular context, such as the emotion labels defined by Feng et al. (2022) to encode task performance simultaneously in ToDs, and (3) depression, a medical illness that negatively affects how a person feels, thinks and acts, and causes feelings of sadness and/or a loss of interest in activities the person once enjoyed (Amer-

ican Psychiatric Association, 2020).

The emergence of LLMs has signified a shift of paradigm from training small models for one specific task to large models for multiple tasks. Therefore, in this work, we investigate the affect recognition ability of a range of LLMs on vastly different types of dialogues and labels<sup>1</sup> to ascertain the validity of this direction. Specifically,

- We evaluated and compared the ability of a range of LLMs to recognise human affect under different dialogue set-ups (chit-chat dialogues and ToDs) and recognition targets (emotion classes and binary depression diagnosis). We used the following datasets: IEMOCAP (Busso et al., 2008), EmoWOZ (Feng et al., 2022), and DAICWOZ (Gratch et al., 2014).
- We investigated into LLMs’ zero-shot and few-shot capabilities through an array of ICL set-ups as well as their model capacities through task-specific fine-tuning.
- We considered text-based LLMs as a part of spoken dialogues systems. Therefore, we also experimented with inputs containing automatic speech recognition (ASR) errors to investigate the potential influence of ASR errors on LLM predictions.

## 2 Related Work

### 2.1 LLM

Large Language Model (LLM) refers to a type of pre-trained models designed for natural language processing tasks. LLMs are characterised by their enormous number of model parameters and extensive training data.

Some well-known examples of LLMs include OpenAI GPT family models (Radford et al., 2019; OpenAI, 2022, 2023), which can have billions or even trillions of model parameters. Examples of open-source text-based foundation models include the LLaMA family (Touvron et al., 2023a,b; AI@Meta, 2024) and their corresponding chat-optimised models. These models have demonstrated remarkable abilities in various natural language understanding and generation tasks, including text completion, language translation, text summarisation, and even chatbot applications (Beeching et al., 2023). They also demonstrate “emergent abilities” such as few-shot prompting and chain-of-thought reasoning, which were not present in their smaller

<sup>1</sup>The code can be found at <https://gitlab.cs.uni-duesseldorf.de/general/dsml/llm4erc-public/>

predecessors (Wei et al., 2022). While there are also multi-modal LLMs such as SALMONN (Tang et al., 2024), these are at an earlier stage compared to uni-modal text-based LLMs, and it is still a common practice to use text-based LLMs as the text-processing backend, pipelined with other modules such as ASR and image generator for more complex applications.

### 2.2 Affective Capabilities of LLMs

With the growing attention on LLMs from the research community, there have been several works investigating the affective abilities of LLMs. Huang et al. (2023) evaluated the empathy ability of LLMs by utilising the emotion appraisal theory from psychology. Wang et al. (2023) assessed the emotional intelligence of LLMs in terms of Emotional Quotient (EQ) scores. Zhang et al. (2023) investigated how LLMs could be leveraged for a range of sentiment analysis tasks under zero-shot or few-shot learning set-ups. Zhao et al. (2023) investigated the emotional dialogue ability of ChatGPT through a range of understanding and generation tasks. In our work, we focus on the affect recognition ability of text-based LLMs. Our investigation spans across different types of dialogues and model learning set-ups. We also consider real-world applications of LLMs and consider ASR-inferred noisy input to models.

## 3 Methodology

The ability of human-beings to recognise affect can be reflected in many ways. Yet, being able to narrate what emotion has been expressed in the utterances of the other interlocutor is a straightforward and strong sign of such an ability. Therefore, we took LLMs’ ability to verbalise the emotion given the dialogue context as a proxy to both qualitatively and quantitatively analyse LLMs’ ability for affect recognition.

### 3.1 Affect Recognition using LLMs

The pipeline for affect recognition using LLMs with the option to take speech as input is illustrated in Figure 1. When using the speech input, a Whisper-medium model was used to transcribe the speech (see Section 4.5 for details). The prompt is then constructed as designed and fed into the LLM to generate a text sequence. For open-source LLMs, we examined the probability of each class token and considered the one with the maximum probability as the final model prediction, as shown

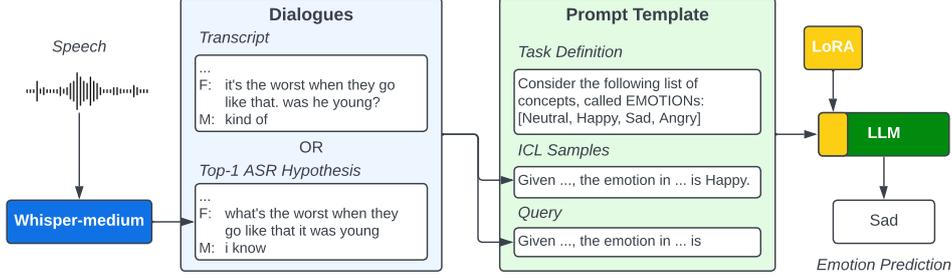


Figure 1: A flowchart illustrating the affect recognition pipeline using Whisper and LLM. The designed prompt comprises parts introduced in Table 1. Low-rank adaptation (LoRA) is used for fine-tuning open-source LLMs.

in Equation 1.

$$\mathbf{W}_{L^*} = \arg \max_{\mathbf{W}_L} P(\mathbf{W}_L | \mathbf{W}^P), \quad (1)$$

where  $\mathbf{W}_L$  belongs to the set of pre-defined labels and  $\mathbf{W}^P$  is the prompt token sequence.

For commercial models, there is no access to logits of model outputs and model outputs do not always follow the format specified in the prompt. Therefore, we used regular expressions to derive the final prediction.

### 3.2 Task-specific Fine-tuning

For efficient training of LLMs, we utilise low-rank adaptation (LoRA, Hu et al. 2022) to accelerate the fine-tuning of LLMs while conserving memory. This is also a common approach for fine-tuning LLMs as seen in many existing works (Sun et al., 2023; Zhao et al., 2024).

LoRA hypothesises that the change in weights during model training has a low “intrinsic rank”. Therefore, instead of directly updating the full-rank weight matrices of dense layers during training, LoRA optimises the low-rank decomposition matrices of those dense layers’ changes while keeping the pre-trained weights frozen. Specifically, for a pre-trained weight matrix  $\mathbf{W}_0 \in \mathbb{R}^{m \times n}$  from a particular attention block in a transformer-based LLM, its update  $\Delta \mathbf{W}$  is constrained using a low-rank decomposition of the update as following:

$$\mathbf{W}_0 + \Delta \mathbf{W} = \mathbf{W} + \mathbf{A}\mathbf{B} \quad (2)$$

where matrices  $\mathbf{A} \in \mathbb{R}^{m \times r}$  and  $\mathbf{B} \in \mathbb{R}^{r \times n}$  contain trainable parameters and  $r \leq \min(m, n)$ . The pre-trained parameters in  $\mathbf{W}_0$  are fixed. When  $r$  is set to a much smaller value than the dimensions of  $\mathbf{W}_0$ , the number of trainable parameters will be greatly reduced. This leads to greater training

efficiency, less memory requirement, and a lower chance of over-fitting. Following Hu et al. (2022), we apply LoRA to the projection matrices of the self-attention layers of transformer-based LLMs.

LLMs are trained to predict the next token in the sequence (the label tokens), given the previous tokens (the designed prompt). During training, the input tokens are fed into the model, and the model predicted the probability distribution of the next token. The cross-entropy loss is calculated from the model prediction and the target token.

With LoRA, it takes roughly 30GB memory and 4 hours to train one epoch on the entire EmoWOZ training set using an Nvidia A100 40GB graphics card.

## 4 Experimental Setup

### 4.1 Datasets and Evaluation

The IEMOCAP dataset (Busso et al., 2008) is a multi-modal corpus designed for Emotion Recognition in Conversations (ERC) task in chit-chat or open-domain dialogues. It comprises 151 dialogues, containing 10,039 utterances from 10 distinct speakers involved in 5 dyadic conversational sessions. Each utterance underwent annotation by a minimum of three annotators, who assigned one of nine emotion classes, including *sad*, *neutral*, *angry*, *happy*, *excited*, *frustrated*, *surprised*, *fearful*, *disgusted*. Annotators could also assign multiple emotions or use the category “other” if the perceived emotion did not match the predefined options. Final labels were determined via majority voting.

Given the absence of an official train-test split, we adopt leave-one-session-out 5-fold cross-validation approach and average the results. Our methodology aligns with the common practices, as discussed

by Wu et al. (2020), to consider two label sets: **4-way**: *Sad, Neutral, Angry, and Happy*; **5-way**: *Sad, Neutral, Angry, Happy, and everything else as Other*. In both set-ups, *Excited* is merged with *Happy*.

Emotion recognition is performed for every speaker utterance. We report the weighted accuracy (WA) and unweighted accuracy (UA) for both label sets.

**EmoWOZ** (Feng et al., 2022) is a text-based ERC corpus built for emotion recognition in ToDs. It comprises 10,438 human-human dialogues from the entire MultiWOZ dataset (Budzianowski et al., 2018), as well as 1,000 human-machine dialogues in the same set of domains. It encompasses seven distinct user emotions, namely: *Neutral, Fearful, Dissatisfied, Apologetic, Abusive, Excited, and Satisfied*. These emotion labels are designed to encode the task performance. Specifically, each emotion is defined as a valence reaction to certain elicitor under certain conduct. For example, *Dissatisfied* is defined as a negative emotion elicited by the system expressed in a neutral or polite conduct.

Emotion recognition is performed for each **user** utterance. For existing benchmarks reported in Feng et al. (2022), neutral class was excluded from calculating the metrics because they take up more than 70% of the labels in EmoWOZ. To have a direct comparison, we report macro-averaged F1 and weighted average F1 excluding neutral. We include the F1, precision, and recall of the neutral class in Table B3 of Appendix B.

**DAIC-WOZ** (Gratch et al., 2014) is a speech-based corpus for depression detection and analysis. It includes the Patient Health Questionnaire-8 (PHQ-8, Kroenke et al., 2008) scores of 193 clinical interviews, with 35 (12 are labelled depressed) interviews in the development set and 47 (14 are labelled depressed) in the test set. The PHQ-8 score ranges from 0 to 24 and quantifies the severity of the patient’s depressive symptoms.

For evaluation metrics, we follow the criteria established by the Audio/Visual Emotion Challenge and Workshop challenge (AVEC2016) (Valstar et al., 2016) and perform binary classification on the dialogue level. Interviewees with  $\text{PHQ8} \geq 10$  is considered *Depressed* and  $\text{PHQ8} < 10$  is considered *Not Depressed*. Since patients with PHQ-8 score of 5 to 9 are defined to show mild depressive symptoms (Kroenke et al., 2008) but considered

*Not Depressed* in the dataset, we add information about PHQ-8 level definition and quantisation criteria to the prompt to establish an aligned diagnosis standard (Table 1) for the model.

Notably, participants in the AVEC2016 challenge (Yang et al., 2016; Williamson et al., 2016) and subsequent research efforts (Ravi et al., 2022; Wu et al., 2023) primarily focused on optimising the F1 score of the *Depressed* class. We report this metric in Section 5 for direct comparison and also include the F1 score of the *Not Depressed* in Appendix B.

## 4.2 Prompt Design

The prompt design aims to exploit the language modelling and in-context learning ability of LLMs. Due to the different task set-ups and label sets in each datasets, the prompt templates used are different as illustrated in Table 1. Specifically, EmoWOZ uses custom emotion labels, DAIC-WOZ involves mapping from numerical values to binary classes, and IEMOCAP uses generic emotion labels. We therefore provide additional label explanations in the task definition of EmoWOZ and DAIC-WOZ. IEMOCAP on the other hand, contains self-explanatory emotion labels from a generic emotion model and does not come with any special definitions. Therefore, we do not include label definition in the prompt for IEMOCAP. Since IEMOCAP and EmoWOZ involve utterance-level classification whereas DAIC-WOZ involves dialogue-level classification, we used different queries to accommodate this difference.

## 4.3 Models

### 4.3.1 LLMs

**GPT-2** (Radford et al., 2019) has a transformer architecture, pretrained on a substantial English corpus through self-supervised learning. While its size does not make it one of LLMs, it stands as one of the early achievements of OpenAI’s GPT models. For our baseline reference, we utilised the version containing 124 million parameters.

**GPT-3.5**, or ChatGPT (OpenAI, 2022), is a chatbot application developed by OpenAI. It follows a similar architecture as InstructGPT (Ouyang et al., 2022) and was fine-tuned for chat application via reinforcement learning from human feedback (RLHF). It contains 175 billion parameters. Specifically, we used the version released on the 13<sup>th</sup> of June, 2023.

Prompt Template	
Task Definition	<b>IEMOCAP:</b> Consider the following list of concepts, called EMOTIONS: [Emotion <sub>A</sub> , Emotion <sub>B</sub> , ...] <b>EmoWOZ:</b> Consider the following list of concepts, called EMOTIONS: [Emotion <sub>A</sub> : Emotion_Definition <sub>A</sub> ; Emotion <sub>B</sub> : Emotion_Definition <sub>B</sub> ; ...] <b>DAIC-WOZ:</b> Given that the SEVERITY of depression can be categorised into the following levels on a scale of 0 to 24: [No significant depressive symptoms (0 to 4), ...]. A participant is considered depressed if the participant shows moderate depressive symptoms (10 to 14) and above.
ICL Samples	<b>IEMOCAP / EmoWOZ:</b> Given the dialogue history between Speaker <sub>A</sub> and Speaker <sub>B</sub> : [Speaker <sub>A</sub> : Utterance <sub>t-3</sub> ; Speaker <sub>B</sub> : Utterance <sub>t-2</sub> ; Speaker <sub>A</sub> : Utterance <sub>t-1</sub> ], the EMOTION in the next utterance "Speaker <sub>B</sub> : Utterance <sub>t</sub> " is Emotion <sub>A</sub> <b>DAIC-WOZ:</b> Given the depression consultation dialogue between Participant and Ellie: [Participant: Utterance <sub>0</sub> ; Ellie: Utterance <sub>1</sub> ; Participant: Utterance <sub>2</sub> ; ...], the Participant's is (not) depressed.
Query	<b>IEMOCAP / EmoWOZ:</b> Given the dialogue history between Speaker <sub>A</sub> and Speaker <sub>B</sub> : [Speaker <sub>A</sub> : Utterance <sub>t-3</sub> ; Speaker <sub>B</sub> : Utterance <sub>t-2</sub> ; Speaker <sub>A</sub> : Utterance <sub>t-1</sub> ], the EMOTION in the next utterance "Speaker <sub>B</sub> : Utterance <sub>t</sub> " is <b>DAIC-WOZ:</b> Given the depression consultation dialogue between Participant and Ellie: [Participant: Utterance <sub>0</sub> ; Ellie: Utterance <sub>1</sub> ; Participant: Utterance <sub>2</sub> ; ...], the Participant's is

Table 1: Prompt templates, consisting of the task definition, in-context samples, and the query.

**GPT-4** (OpenAI, 2023) is an improved version of GPT-3.5. Its size is six times that of GPT-3.5. Although it is considered a multi-modal model because it additionally accepts images as input, we only explored its text modality. We used the version released on the 13<sup>th</sup> of June, 2023.

**LLaMA-7B** (Touvron et al., 2023a) is a large and causal language model introduced by Meta AI in 2023. It has transformer decoder architecture, 7 billion parameters and was pre-trained on 1 trillion tokens.

**Alpaca-7B** (Taori et al., 2023) is fine-tuned from LLaMA-7B with 52K instruction-following demonstrations generated in the style of self-instruct using text-davinci-003, a specific version of Instruct-GPT (Ouyang et al., 2022).

**LLaMA-3-8B** (AI@Meta, 2024) is the most recent model of the LLaMA family, featuring enhanced usefulness and safety. It was pre-trained on 15 trillion tokens.

### 4.3.2 Supervised Models for Comparison

While comparing zero-shot and few-shot ICL results of LLMs with supervised SOTAs does not paint the fairest picture, it does provide us with insights into how far LLMs are from achieving the performance levels of supervised SOTAs.

We compare LLMs' performance with the following supervised models on each dataset: Wu et al. (2020) for IEMOCAP, Feng et al. (2023) for EmoWOZ, and Wu et al. (2023) for DAIC-WOZ. Specifically,

**For IEMOCAP:** Wu et al. (2020) proposed an emotion recognition model which takes 1) a time-synchronous representation that fuses the audio features with the corresponding text information at

each time step, as well as 2) a time-asynchronous representation that captures the text information embedded across the transcriptions of a number of consecutive utterances. These two types of frame-level vectors, after being pooled in their respective branches with self-attentive layers across the input time window, are fused using an fully connected layer for emotion classification.

**For EmoWOZ:** Feng et al. (2023) proposed a model that is dedicated for textual emotion recognition in task-oriented dialogues. Based on a transformer-based classifier that considers the dialogue history and speaker roles, the proposed model adopts data augmentation with chit-chat dialogues, dialogue state features, multi-task classification for emotional aspects, and a distance-based loss that considers the similarity of the custom emotion labels in EmoWOZ.

**For DAIC-WOZ:** Wu et al. (2023) proposed to extract utterance-level representations from pre-trained speech-based foundation model. The foundation model was further fine-tuned for speech recognition and emotion recognition. The average-pooled dialogue-level features were fed into a depression detection block for binary classification. To address the issue of data sparsity in speech depression detection, authors also performed data augmentation using sub-dialogue shuffling.

## 4.4 Training Configurations

We implement LoRA (Section 3.2) when training LLaMA-7B, Alpaca-7B, and LLaMA-3-8B but not GPT-2. For all open-source LLMs, we constrain the decoding space of the model output to ensure it generates the desired class labels. Details can be found in Appendix A.

## 4.5 ASR System Specifications

In order to observe how LLMs perform with the presence of substantial ASR errors rather than building a pipeline for speech-based ERC, we use an “off-the-shelf” OpenAI Whisper-medium model (Radford et al., 2022), which has been trained solely on English data and not been fine-tuned. We use a decoding beam size of 3. The text normalisation only involves removing punctuation marks. The ASR word error rates (WER) for IEMOCAP and DAIC-WOZ are 12.0% and 16.5% respectively. Since EmoWOZ does not come with raw audio data, we build an ASR simulator. We formulate the simulation as a sequence generation task where the source is the ground-truth text and the target is the ASR-transcribed text (as described in Appendix A.2). The resulted simulated WER in EmoWOZ is 17.1%.

## 5 Results and Discussions

In this section, we aim to answer the questions below. Full results can be found in Appendix B.

- How do LLMs perform under zero-shot set-up on different types of dialogues? How robust are LLMs to ASR errors?
- To what extent can few-shot in-context learning improve LLMs’ performance?
- For open-source LLMs, can task-specific fine-tuning achieves SOTA performance on each respective dataset?

### 5.1 Zero-shot Learning

Table 2 summarises LLMs’ zero-shot affect recognition performances on the three datasets, and we made the following observations:

**LLMs’ performance falls short of supervised SOTAs in affect recognition tasks.** Notable gaps are observed when compared the performance achieved by LLMs and supervised SOTAs for all datasets.

It’s noteworthy that although GPT-4, the largest model, underperforms when compared with the supervised SOTA on EmoWOZ, its reported macro-averaged F1 is still comparable to some supervised learning models benchmarked in Feng et al. (2022). This suggests the good capability of GPT-4 in leveraging the label definitions in the prompt to recognise emotions in EmoWOZ, irrespective of their prevalence. Supervised models, however, may be more susceptible to issues such as label imbalance.

**Larger models do not necessarily lead to better performance.** For IEMOCAP, Alpaca-7B demonstrates the best performance, even surpassing much larger models (GPT-3.5 and GPT-4). Conversely, for EmoWOZ and DAIC-WOZ, the performance generally improves as the model size increased.

While chit-chat utterances in IEMOCAP are labelled with emotion classes from generic emotion models, EmoWOZ’s labels are specifically designed to encode the eliciting conditions of emotions in ToDs. This design necessitates more explicit reasoning in ERC within EmoWOZ compared to IEMOCAP. Although LLMs rely on their language modelling capabilities when performing zero-shot ERC, the greater reasoning ability facilitated by the substantial number of parameters in GPT-3.5 and GPT-4 results in improved performance in EmoWOZ.

Likewise in DAIC-WOZ, the recognition is performed for the entire dialogue. Larger models demonstrate greater ability to leverage the more nuanced affective state of the patient in the larger context.

**Fine-tuning LLMs with instruction-following demonstrations facilitates more effective utilisation of the prompt.** In all datasets, Alpaca-7B consistently outperforms LLaMA-7B and even the much more recent LLaMA-3-8B. This indicates that the additional fine-tuning of LLaMA-7B with instruction-following demonstrations has enhanced its capability in ERC.

LLaMA-7B appears to underperform compared to the much smaller GPT-2 on EmoWOZ. This discrepancy can be explained by LLaMA-7B’s strong inclination towards predicting the neutral emotion ( $F1 = 82.1$  with  $Recall = 100$ ), which has been excluded from the metric calculation, resulting in the poor reported metrics. Fine-tuning with instruction-following demonstrations, as adopted in Alpaca-7B, effectively leverages the task and label definition in the prompt and reverts this trend. Such an inclination in predicting neutral emotion in LLaMA-7B does not appear in the more recent LLaMA-3-8B.

### 5.2 Zero-shot Learning with Noisy ASR Input

Table 3 provides a summary of LLMs’ zero-shot performance when replacing the original dialogue transcripts with ASR-inferred inputs. ASR errors

Model	IEMOCAP (4-way)		IEMOCAP (5-way)		EmoWOZ		DAIC-WOZ	
	WA (↑)	UA (↑)	WA (↑)	UA (↑)	MF1 (↑)	WF1 (↑)	F1 (dev, ↑)	F1 (test, ↑)
GPT-2	25.8	29.2	19.0	22.3	7.3	24.0	0.0	0.0
LLaMA-7B	41.1	40.5	35.6	33.6	1.1	0.3	47.5	52.2
Alpaca-7B	<b>48.8</b>	<b>51.4</b>	<b>40.5</b>	<b>36.2</b>	24.0	44.6	47.5	53.3
LLaMA-3-8B	41.8	42.5	29.4	31.7	19.7	42.4	47.1	43.2
GPT-3.5	42.2	37.6	37.9	35.1	39.0	40.0	54.5	<b>64.3</b>
GPT-4	42.4	37.6	37.5	34.7	<b>52.4</b>	<b>62.3</b>	<b>63.6</b>	59.3
Supervised SOTA	77.6	78.4	73.3	74.4	65.9	83.9	88.6	85.7

Table 2: Zero-shot performance of LLMs compared with respective supervised SOTAs. The best zero-shot performance for each metric is made bold. For metrics: WA = weighted average; UA = unweighted average; MF1 = macro-averaged F1 excluding neutral; WF1 = weighted average F1 excluding neutral; F1 = F1 for class *Depressed*.

Model	IEMOCAP (4-way)		IEMOCAP (5-way)		EmoWOZ		DAIC-WOZ	
	WA (↑)	UA (↑)	WA (↑)	UA (↑)	MF1 (↑)	WF1 (↑)	F1 (dev, ↑)	F1 (test, ↑)
LLaMA-7B	-0.3	-1.2	-1.1	-5.0	-1.1	-0.3	-1.6	-1.1
Alpaca-7B	-1.3	-1.8	-1.8	-2.6	+0.3	-2.0	-1.6	+0.0
LLaMA-3-8B	-2.1	-3.5	-1.2	-2.2	+0.1	-0.1	-0.7	-0.3
GPT-3.5	+0.1	-0.1	+0.2	0.0	+1.2	-0.2	-17.0	-8.3
GPT-4	-0.5	-0.5	-1.1	-0.7	+0.9	-1.5	-19.2	-17.6
Supervised SOTA	-3.8	-3.7	-3.9	-3.5	-0.8	-0.4	-3.6	-4.1

Table 3: Change in zero-shot performance metrics of LLMs after using noisy ASR input. For metrics: WA = weighted average; UA = unweighted average; F1 = F1 for class *Depressed*. GPT-2 was omitted due to its poor zero-shot capability.

exhibit varying degrees of influence on different affect recognition tasks. Specifically,

**LLMs are generally robust to ASR errors when recognising emotion.** This is exemplified by small changes in metrics for IEMOCAP compared with supervised SOTAs. The only one notable exception is the UA of LLaMA-7B in the 5-way classification task on IEMOCAP. Looking at the performance of each emotion in this experiment, we observed significant drops in the F1 scores for the emotions *{Happy, Angry, and Sad}*. Specifically, *Happy* and *Angry* experience major decreases in their recall scores (*Happy*: 12.3  $\rightarrow$  7.3, *Angry*: 50.0  $\rightarrow$  11.0), while *Sad* sees a substantial decline in its precision score (65.5  $\rightarrow$  0.0). At the same time, there is an increase in the recall score for the *Other* category (47.3  $\rightarrow$  78.2), resulting in an overall rise in its F1 score (44.5  $\rightarrow$  48.0). These observations suggest that ASR errors introduced a tendency for LLaMA-7B to mis-classify more emotions as *Other*.

**ASR errors have a more pronounced influence on the accuracy of depression detection.** For DAIC-WOZ, the introduction of ASR errors had a significant impact on F1 scores. The impact diverges for open-source and commercial models.

For open-source models, which are also relatively smaller, the change in F1 was small, showing a similar trend when they recognise emotions from noisy dialogues. On the other hand, for larger commercial models, the F1 scores decrease more significantly. This phenomenon can be ascribed to the lengthy prompt for conducting dialogue-level analysis, in which ASR errors accumulated. While OpenAI models can better leverage information from the large context, such an ability adversely affects its depression detection ability in the presence of ASR errors.

### 5.3 In-context Learning

ICL samples are randomly selected for each class and are the same within each experiment set-up for all models. The performance of LLMs with different numbers of ICL samples is outlined in Table 4, from which we have derived the following observation:

**Larger models tend to derive greater benefits from an increased number of ICL samples to recognise emotions.** LLaMA-7B, Alpaca-7B, and LLaMA-3-8B do not consistently benefit from an increased number of ICL samples in the prompt. Optimal model performance generally occurs when  $N = 0$  or  $N = 1$ . This suggests that effectively

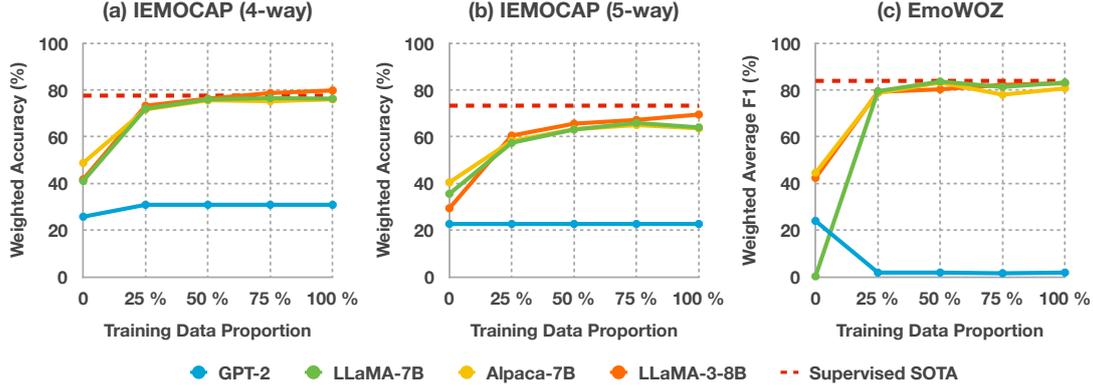


Figure 2: Change of model performance when fine-tuning with different proportions of the training data.

Model	N	IEMOCAP		EmoWOZ	DAIC-WOZ	
		4-way	5-way		Dev	Test
LLaMA-7B	0	41.1	<b>35.6</b>	0.3	<b>47.5</b>	<b>52.2</b>
	1	<b>52.3</b>	27.3	<b>42.6</b>	0.0	0.0
	3	42.8	26.2	27.2	42.1	48.9
Alpaca-7B	0	48.8	<b>40.5</b>	44.6	<b>47.5</b>	<b>53.3</b>
	1	<b>54.1</b>	26.9	<b>51.2</b>	0.0	15.4
	3	52.4	24.4	44.6	45.9	51.1
LLaMA-3-8B	0	41.8	29.4	<b>42.4</b>	<b>47.1</b>	<b>43.2</b>
	1	56.8	<b>40.5</b>	38.0	0.0	0.0
	3	<b>57.4</b>	24.4	39.9	0.0	0.0
GPT-3.5	0	42.2	37.9	40.0	<b>54.5</b>	<b>64.3</b>
	1	56.3	<b>48.3</b>	43.2	13.3	40.0
	3	<b>62.1</b>	<b>48.3</b>	<b>46.7</b>	37.5	56.0
GPT-4	0	42.4	37.5	62.3	63.6	<b>59.3</b>
	1	62.9	49.0	64.4	<b>80.0</b>	55.6
	3	<b>63.8</b>	<b>49.4</b>	<b>66.5</b>	74.1	58.5

Table 4: Performance of LLMs (WA for IEMOCAP and WF1 for EmoWOZ) under in-context learning set-ups. N stands for the number of ICL samples per emotion class and  $N = 0$  means the zero-shot set-up. The best performance of each model is made bold.

utilising the full context remains as a challenge for LLMs. Larger models, GPT-3.5 and GPT-4, show more consistent improvement in performance with the increased number of ICL samples. GPT-4 derives the most significant benefits from ICL samples and performs the best across all models.

**The effectiveness of ICL is limited for depression detection.** The performance is in general the best when  $N = 0$ , followed by  $N = 3$ . This suggests that for depression detection, a task to detect more nuanced affective state than emotion from a longer sequence, a single ICL sample for each class could strongly bias the model. This leads to zero F1s where models predict all samples as *Not Depressive*. Including more ICL samples could mitigate this effect, but the performance is further limited by models’ incapability to handle extremely lengthy input. This motivates further research ef-

forts to handle huge context containing nuanced task-related cues when using LLMs.

#### 5.4 Task-specific Fine-tuning

We conduct task-specific fine-tuning experiments with GPT-2, LLaMA-7B, Alpaca-7B, and LLaMA-3-8B using different proportions of training data to explore these models’ capacity for ERC after fine-tuning. Results are summarised in Figure 2. For DAIC-WOZ, fine-tuning would steer models to predict *Not Depressed* (see Table B4) for almost all test samples. This might be due to the small training set where more than 70% of the samples are labelled as *Not Depressed*. This suggests the limitation of language modelling objective, and therefore more carefully curated task-related learning objectives should be considered for depression detection using LLM.

#### Task-specific fine-tuning can effectively and efficiently enhance the ERC performance of LLMs.

For both IEMOCAP and EmoWOZ, we observe an initial significant improvement in performance when fine-tuning with 25% of the training data. Performance remains relatively stable and approaches SOTA levels as the proportion of training data increased to 50% and more for IEMOCAP (4-way) and EmoWOZ. This shows the potential of rapid deployment of LLMs as the emotion recognition frontend in dialogue systems, regardless of dialogue type, label set, or label distribution.

In the case of 5-way classification on IEMOCAP, a performance gap persists between fine-tuned LLMs and the supervised SOTA, even after fine-tuning of LLMs on the complete training set. We hypothesised that this disparity might be attributed to the presence of an additional *Other* class within

the 5-way classification scheme. The class name “Other” lacked essential affective information and consequently failed to fully leverage the language modelling capabilities of LLMs. Therefore, we suggest that employing more semantically meaningful label names could be advantageous in harnessing the potential of LLMs for task-specific fine-tuning.

In the case of GPT-2, fine-tuning does not yield noticeable improvement in ERC. Its performance even deteriorated after fine-tuning with EmoWOZ, as depicted in Figure 2(c) because GPT-2 predominantly predicted *Neutral*, which are excluded from the metric calculation.

## 6 Conclusion

In this study, we explore the performance of LLMs for affect recognition in three distinct types of dialogues: chit-chat dialogues, information-seeking ToDs, and medical consultation dialogues for depression. We conduct benchmark experiments on these datasets using five LLMs: LLaMA-7B, Alpaca-7B, LLaMA-3-8B, GPT-3.5, and GPT-4. We also explore various setups, including zero-shot learning, few-shot in-context learning, and task-specific fine-tuning, all facilitated by specially designed prompts. Additionally, we examine the impact of ASR errors on LLMs’ zero-shot performance.

Our zero-shot experiments underscore that while LLMs have made significant strides in various natural language understanding tasks, they still have some distance to cover in order to match the supervised SOTAs in affect recognition tasks. Adding emotion definitions explaining the eliciting conditions in ToDs to the prompt and fine-tuning LLMs for instruction-following could narrow the performance gap from supervised SOTAs.

Performing zero-shot affect recognition from utterances containing ASR errors shows that LLMs are robust to such errors for emotion recognition but not for depression detection. Therefore, when considering LLMs as a back-end module of a spoken dialogue system, it is crucial to exercise extra caution when processing dialogues laden with highly specific and nuanced affective content.

Our ICL experiments exemplify that larger models would benefit more from an increased number of ICL samples, highlighting the need to explore the optimal combination of the ICL sample size in the

prompt and the model size. For smaller LLMs, effectively utilising lengthy context remains as a challenge.

Through task-specific fine-tuning, we achieve performance levels close to SOTA on IEMOCAP and EmoWOZ, using only 50% of the training data, with LLaMA-7B, Alpaca-7B, and LLaMA-3-8B. This highlights the great potential of fine-tuning LLMs for simpler tasks and integrating them as functional modules into dialogue systems.

Overall, LLMs have opened new avenues for affect recognition in conversations and building affect-aware dialogue systems. Despite the limited performance under zero-shot set-up, their robustness to ASR errors, few-shot ICL capabilities, and ERC capabilities after fine-tuning offer exciting research opportunities for exploring affect recognition in conversations and building human-like conversational agents. We would also like to highlight the challenge and also opportunities towards handling long context and nuanced emotion cues in LLMs.

## 7 Limitations

In our work, although we reduce computation resource of training LLMs by incorporating LoRA, the inference takes  $\tilde{t}$ s for utterance-level emotion recognition on a Nvidia A100 40GB graphics card when there is no ICL sample in the prompt. The inference time increases when the number of ICL samples increases or dialogue-level classification is performed. While LLMs demonstrates superior abilities and potentials, further research efforts are still needed to ensure efficient LLM inference, which is necessary for its application in real-time systems.

With ICL experiments especially on DAIC-WOZ, we observe that the efficacy of long context is limited by the effective spans of the attention mechanisms. While substantial efforts have been invested into increasing the maximum allowed context size of LLMs and improving benchmark performance, the effectiveness of LLMs to make use of full context should not be overlooked.

We only investigate with one dataset from each of three dialogue domains. Although these datasets cover different dialogue settings, objectives, label sets, and classification scopes, there are more affect types and dialogue settings to explore. These datasets also exhibit various degrees of class im-

balance, which selected reference SOTAs utilised data augmentation to address. While GPT-4 has demonstrated good zero-shot learning ability (Section 5.1), addressing data imbalance is out of the scope of this work, and data augmentation with LLMs may come at a cost of potential divergence between synthetic language and real-word data (Li et al., 2023).

## 8 Ethics Statement

Models and datasets were used in accordance with their respective licenses. Data that we used and generated does not contain any information that uniquely identifies individual people. There is a tiny fraction of utterances labelled as “abusive” in EmoWOZ, but they are prompted to models in such a way for the recognition purpose only. Due to the fact that LLMs were pre-trained with a huge amount of data, they may produce inaccurate information about people, places, or facts. This had negligible impact on our evaluation for affect recognition. When performing depression detection and analysis with DAIC-WOZ using GPT-3.5 and GPT-4, models output reminders about seeking professional advice from doctors for more accurate medical diagnosis along with their predictions.

Unlike running models locally, utilising OpenAI’s server-based models would require us to send data to their server. In some cases, it is important to use the application programming interface (API) when for which OpenAI explicitly clarifies that the query data will not be stored or used in model training unless specifically configured.

Although this work focuses on LLMs’ capability in recognising affect in conversations, we do envisage LLMs to be incorporated as an affect recognition frontend in affect-aware dialogue systems. It is therefore important to remember that these models are not perfect and can make errors in their predictions. Subsequently, any actions taken based on these predictions should be executed with an awareness of the possibility of errors. The relatively slow inference speed and the high computational resource requirement also pose a challenge in the usage of LLMs in high-throughput and time-sensitive scenarios.

## 9 Acknowledgement

S. Feng and N. Lubis are supported by funding provided by the Alexander von Humboldt Founda-

tion in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research. G. Sun is partly funded by the Department of Engineering, University of Cambridge. Computing resources were provided by Google Cloud.

## References

- AI@Meta. 2024. *Llama 3 model card*.
- American Psychiatric Association. 2020. What Is Depression? <https://www.psychiatry.org/patients-families/depression/what-is-depression>.
- Edward Beeching, Clémentine Fourier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard).
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. *MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.
- Paul Ekman and W V Friesen. 1971. *Constants across cultures in the face and emotion*. *Journal of personality and social psychology*, 17 2:124–9.
- Shutong Feng, Nurul Lubis, Christian Geishhauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. *EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.
- Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishhauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gasic. 2023. *From chatter to matter: Addressing critical steps of emotion recognition learning in task-oriented dialogue*. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 85–103, Prague, Czechia. Association for Computational Linguistics.
- Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Strattou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. *The*

- distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).
- James J. Gross. 2002. Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3):281–291.
- Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geischauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Jen-tse Huang, Man Lam, Eric Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench.
- Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.
- Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet B W Williams, Joyce T Berry, and Ali H Mokdad. 2008. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*, 114(1-3):163–173.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>.
- John D Mayer, David R Caruso, and Peter Salovey. 1999. Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27(4):267–298.
- OpenAI. 2022. Introducing ChatGPT. <https://openai.com/blog/chatgpt>.
- OpenAI. 2023. Gpt-4 technical report.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS 2022*.
- Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. 2022. A step towards preserving speakers’ identity while detecting depression via speaker disentanglement. *Interspeech*, 2022:3338–3342.
- James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Guangzhi Sun, Shutong Feng, Dongcheng Jiang, Chao Zhang, Milica Gašić, and Philip C. Woodland. 2023. Speech-based slot filling using large language models.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutit Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams,

Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. [Avec 2016: Depression, mood, and emotion recognition workshop and challenge](#). In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 3–10, New York, NY, USA. Association for Computing Machinery.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023. [Emotional intelligence of large language models](#).

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*. Survey Certification.

James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrienne Schwarzentruher, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F. Quatieri. 2016. [Detecting depression using vocal, facial and semantic communication cues](#). In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 11–18, New York, NY, USA. Association for Computing Machinery.

Wen Wu, Chao Zhang, and Philip C. Woodland. 2020. [Emotion recognition by fusing time synchronous and time asynchronous representations](#). *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6269–6273.

Wen Wu, Chao Zhang, and Philip C. Woodland. 2023. [Self-supervised representations in speech-based depression detection](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2016. [Decision tree based depression classification from audio video and language information](#). In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 89–96, New York, NY, USA. Association for Computing Machinery.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. [A survey of affect recognition methods: Audio, visual, and spontaneous expressions](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. [Sentiment analysis in the era of large language models: A reality check](#).

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. [Lora land: 310 fine-tuned llms that rival gpt-4, a technical report](#).

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. [Is chatgpt equipped with emotional dialogue capabilities?](#)

## A Detailed Training Configurations

### A.1 Task-Specific Fine-tuning

For all model fine-tuning, the learning rate was  $3e-5$ . The batch size was 2 with a gradient accumulation step of 4. We used a sinusoidal learning rate scheduler without warming up. We applied a weight decay of 0.01 on all model parameters except for the biases and layer normalisation weights. For LLaMA-7B, Alpaca-7B, and LLaMA-3-8B, we stored model parameters in IEEE 754 half-precision float point format. For GPT-2, we stored the model parameters in standard single-precision floating-point format and did not apply LoRA during the fine-tuning. We followed the default LoRA configuration provided in Huggingface PEFT library (Mangrulkar et al., 2022). We used the model perplexity on the development set as the early-stopping criterion. For EmoWOZ, we used the official development set. For IEMOCAP, when we performed the leave-one-session-out training, 10% of the training data were randomly sampled as the development set. We applied stratified sampling based on the emotion labels. All open-source models were trained on a single Nvidia A100 40GB Graphics Card.

### A.2 ASR Simulation for EmoWOZ

We fine-tuned a LLaMA-7B model using LoRA following configurations specified in Section 3.2 and A.1 for one epoch on all IEMOCAP utterances. The source was each of the IEMOCAP utterance transcription and the target was the corresponding OpenAI Whisper-medium hypothesis. We utilised a prompt template that formatted the source and target in natural language would best exploit the language modelling capability of the model:

After adding automatic speech recognition errors, [SOURCE] becomes [TARGET]

## B Detailed Experimental Results

Model	N	P	M	Neutral	Happy	Angry	Sad	WA	UA
GPT-2	0	0%	T	0.7 (60.0/0.4)	32.3 (43.6/25.6)	35.3 (22.0/90.6)	0.5 (30.0/0.3)	25.8	29.2
GPT-2	1	0%	T	10.9 (43.4/6.2)	9.1 (62.0/4.9)	29.0 (21.8/43.6)	33.3 (22.8/61.9)	24.2	29.2
GPT-2	0	25%	T	47.2 (30.9/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	30.9	25.0
GPT-2	0	50%	T	47.2 (30.9/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	30.9	25.0
GPT-2	0	75%	T	47.2 (30.9/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	30.9	25.0
GPT-2	0	100%	T	47.2 (30.9/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	30.9	25.0
LLaMA-7B	0	0%	T	48.6 (37.5/69.3)	21.8 (82.3/12.5)	53.3 (40.8/76.8)	6.9 (78.0/3.6)	41.1	40.5
LLaMA-7B	0	0%	A	50.3 (37.0/78.8)	14.8 (79.3/8.2)	54.7 (44.9/70.0)	0.6 (100.0/0.3)	40.8	39.3
LLaMA-7B	1	0%	T	55.3 (42.9/77.5)	56.2 (77.3/44.2)	62.0 (55.0/71.0)	11.1 (73.0/6.0)	52.3	49.7
LLaMA-7B	3	0%	T	54.2 (39.4/86.7)	1.2 (90.9/0.6)	44.1 (87.6/29.5)	44.5 (39.5/50.8)	42.8	41.9
LLaMA-7B	0	25%	T	65.1 (64.6/65.6)	77.0 (80.8/73.5)	73.5 (72.6/74.4)	74.2 (71.3/77.3)	72.0	72.7
LLaMA-7B	0	50%	T	69.1 (69.9/68.3)	80.7 (80.6/80.7)	77.3 (78.7/76.0)	78.1 (75.6/80.8)	76.0	76.4
LLaMA-7B	0	75%	T	70.7 (67.3/74.5)	82.2 (84.6/80.0)	76.3 (80.5/72.4)	78.1 (77.5/78.6)	76.5	76.4
LLaMA-7B	0	100%	T	69.7 (66.2/73.5)	82.0 (82.4/81.7)	79.0 (81.0/77.1)	75.8 (80.6/71.6)	76.3	76.0
Alpaca-7B	0	0%	T	34.4 (49.6/26.3)	62.8 (79.3/52.0)	50.2 (34.2/94.6)	44.5 (70.3/32.6)	48.8	51.4
Alpaca-7B	0	0%	A	37.0 (52.5/28.6)	60.7 (75.3/50.8)	48.5 (32.8/93.3)	38.8 (77.6/25.8)	47.5	49.6
Alpaca-7B	1	0%	T	53.8 (49.4/59.0)	59.5 (83.1/46.3)	57.9 (43.5/86.9)	37.0 (72.3/24.8)	54.1	54.3
Alpaca-7B	3	0%	T	55.4 (43.9/74.8)	28.0 (90.6/16.6)	64.5 (60.8/68.6)	54.9 (55.1/54.7)	52.4	53.7
Alpaca-7B	0	25%	T	65.1 (67.3/63.0)	77.0 (77.8/76.2)	74.6 (72.1/77.2)	70.8 (68.8/73.0)	71.7	72.4
Alpaca-7B	0	50%	T	69.1 (69.8/68.3)	80.5 (78.7/82.4)	78.3 (79.7/77.1)	75.7 (75.9/75.6)	75.6	75.8
Alpaca-7B	0	75%	T	70.5 (66.3/75.4)	80.6 (85.2/76.4)	76.3 (78.7/74.2)	74.2 (74.2/74.2)	75.2	75.0
Alpaca-7B	0	100%	T	69.3 (69.1/69.5)	81.0 (82.3/79.8)	78.8 (79.2/78.4)	76.5 (74.7/78.3)	76.0	76.5
LLaMA-3-8B	0	0%	T	3.4 (55.6/1.8)	55.9 (42.7/81.0)	51.0 (38.7/75.0)	19.9 (55.7/12.1)	41.8	42.5
LLaMA-3-8B	0	0%	A	2.1 (56.2/1.1)	51.2 (36.1/88.3)	52.3 (47.4/58.5)	14.3 (65.4/8.0)	39.7	39.0
LLaMA-3-8B	1	0%	T	52.7 (50.1/55.6)	67.1 (75.9/60.1)	60.4 (46.7/85.4)	37.9 (82.9/24.5)	56.8	56.4
LLaMA-3-8B	3	0%	T	35.8 (62.0/25.2)	67.0 (79.5/57.9)	63.4 (53.3/78.2)	60.0 (46.1/86.2)	57.4	61.9
LLaMA-3-8B	0	25%	T	68.2 (66.8/69.6)	78.5 (76.6/80.5)	74.4 (76.0/72.9)	72.4 (76.4/68.7)	73.3	72.9
LLaMA-3-8B	0	50%	T	69.7 (71.0/68.5)	81.4 (79.0/84.0)	77.6 (82.3/73.3)	77.4 (74.8/80.2)	76.3	76.5
LLaMA-3-8B	0	75%	T	71.8 (73.8/70.0)	83.2 (81.7/84.7)	80.6 (82.2/79.1)	80.2 (77.8/82.7)	78.7	79.1
LLaMA-3-8B	0	100%	T	73.2 (74.0/72.5)	84.3 (83.0/85.8)	81.6 (83.1/80.2)	81.0 (80.4/81.6)	79.8	80.0
GPT-3.5	0	0%	T	51.6 (35.1/97.3)	28.5 (90.5/16.9)	31.7 (79.9/19.8)	27.1 (81.5/16.2)	42.2	37.6
GPT-3.5	0	0%	A	51.3 (34.9/96.5)	31.7 (88.8/19.3)	33.0 (83.2/20.6)	23.2 (80.8/13.6)	42.3	37.5
GPT-3.5	1	0%	T	57.7 (42.7/88.9)	59.4 (84.8/45.7)	56.6 (76.1/45.1)	45.9 (80.0/32.2)	56.3	53.0
GPT-3.5	3	0%	T	60.1 (48.9/78.2)	66.0 (80.0/56.1)	63.8 (76.4/54.8)	59.4 (67.1/53.3)	62.1	60.6
GPT-4	0	0%	T	51.7 (35.0/99.3)	28.9 (97.5/16.9)	29.4 (95.5/17.4)	28.2 (91.4/16.7)	42.4	37.6
GPT-4	0	0%	A	51.5 (34.8/98.9)	27.6 (94.3/16.2)	30.4 (95.7/18.0)	25.9 (89.6/15.1)	41.9	37.1
GPT-4	1	0%	T	62.3 (48.1/88.2)	59.7 (83.4/46.5)	70.4 (81.6/61.9)	60.9 (81.1/48.7)	62.9	61.3
GPT-4	3	0%	T	61.6 (49.0/83.1)	60.7 (84.9/47.2)	67.6 (85.2/56.1)	68.4 (71.6/65.4)	63.8	63.0

Table B1: F1(precision/recall), UA, and WA of LLMs on IEMOCAP under the 4-Way classification set-up. In table headers, “N” stands for the number of ICL samples in the prompt; “P” stands for the proportion of training data used for fine-tuning; “M” stands for the modality of input, either transcription (T) or ASR hypothesis (A).

Model	N	P	M	Neutral	Happy	Angry	Sad	Other	WA	UA
GPT-2	0	0%	T	0.1 (7.7/0.1)	23.1 (35.2/17.2)	27.5 (16.5/81.1)	0.5 (42.9/0.3)	15.2 (19.3/12.5)	19.0	22.3
GPT-2	1	0%	T	13.3 (31.6/8.4)	26.6 (35.4/21.3)	19.2 (16.3/23.5)	27.4 (17.0/70.5)	0.0 (0.0/0.0)	20.1	24.7
GPT-2	0	25%	T	37.0 (22.7/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	22.7	20.0
GPT-2	0	50%	T	37.0 (22.7/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	22.7	20.0
GPT-2	0	75%	T	37.0 (22.7/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	22.7	20.0
GPT-2	0	100%	T	37.0 (22.7/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	22.7	20.0
LLaMA-7B	0	0%	T	38.7 (29.8/55.2)	21.5 (81.8/12.3)	37.8 (30.3/50.0)	6.3 (65.5/3.3)	44.5 (42.1/47.3)	35.6	33.6
LLaMA-7B	0	0%	A	37.0 (30.9/46.3)	13.4 (80.5/7.3)	17.3 (40.5/11.0)	0.0 (0.0/0.0)	48.0 (34.6/78.2)	34.5	28.6
LLaMA-7B	1	0%	T	2.2 (59.4/1.1)	0.1 (100.0/0.1)	15.1 (60.9/8.6)	0.0 (0.0/0.0)	41.6 (26.5/97.2)	27.3	21.4
LLaMA-7B	3	0%	T	11.1 (22.4/7.4)	0.2 (16.7/0.1)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	41.2 (26.5/92.3)	26.2	20.0
LLaMA-7B	0	25%	T	51.2 (48.1/54.9)	74.3 (72.4/76.3)	49.3 (54.6/44.9)	62.5 (63.3/61.6)	49.9 (51.3/48.6)	57.4	57.3
LLaMA-7B	0	50%	T	57.9 (54.5/61.7)	79.0 (77.1/80.9)	54.7 (57.7/52.0)	69.3 (73.8/65.3)	55.6 (56.7/54.6)	63.1	62.9
LLaMA-7B	0	75%	T	60.8 (58.2/63.7)	82.3 (82.2/82.5)	56.4 (60.9/52.5)	72.9 (68.9/77.4)	57.3 (59.6/55.2)	65.9	66.3
LLaMA-7B	0	100%	T	53.9 (57.5/50.6)	80.9 (77.6/84.5)	55.8 (63.6/49.8)	72.1 (69.8/74.5)	57.4 (54.5/60.7)	64.0	64.0
Alpaca-7B	0	0%	T	18.1 (42.5/11.5)	52.6 (78.3/39.6)	29.2 (34.9/25.0)	29.2 (69.4/18.5)	48.4 (33.6/86.6)	40.5	36.2
Alpaca-7B	0	0%	A	15.8 (41.3/9.8)	49.5 (74.3/37.1)	23.4 (31.5/18.6)	21.8 (72.8/12.8)	48.2 (32.9/90.0)	38.7	33.6
Alpaca-7B	1	0%	T	0.2 (28.6/0.1)	0.5 (100.0/0.2)	5.1 (74.4/2.6)	0.0 (0.0/0.0)	42.0 (26.6/99.5)	26.9	20.5
Alpaca-7B	3	0%	T	9.4 (17.1/6.4)	3.7 (24.3/2.0)	3.1 (18.1/1.7)	5.7 (10.4/3.9)	39.6 (26.1/81.5)	24.4	19.1
Alpaca-7B	0	25%	T	48.9 (53.2/45.3)	73.8 (68.0/80.8)	51.2 (56.4/46.9)	63.4 (60.5/66.5)	52.1 (51.8/52.4)	58.2	58.4
Alpaca-7B	0	50%	T	56.4 (54.0/59.1)	78.7 (76.8/80.6)	55.5 (58.9/52.4)	68.8 (74.6/63.8)	57.4 (57.0/57.9)	63.2	62.8
Alpaca-7B	0	75%	T	57.9 (59.5/56.4)	81.0 (77.5/84.9)	58.9 (62.5/55.7)	70.2 (64.8/76.7)	56.9 (59.2/54.9)	65.0	65.7
Alpaca-7B	0	100%	T	54.7 (55.6/53.8)	81.1 (77.7/84.8)	58.0 (60.3/55.9)	69.8 (65.3/75.0)	55.1 (57.8/52.6)	63.6	64.4
LLaMA-3-8B	0	0%	T	1.6 (46.7/0.8)	44.5 (30.1/85.3)	38.5 (27.3/65.1)	12.4 (40.9/7.3)	0.4 (10.3/0.2)	29.4	31.7
LLaMA-3-8B	0	0%	A	0.9 (44.4/0.5)	41.0 (26.5/90.7)	39.5 (32.4/50.6)	9.2 (47.8/5.1)	1.6 (23.6/0.8)	28.2	29.5
LLaMA-3-8B	1	0%	T	44.1 (43.5/44.8)	68.5 (66.7/70.4)	39.8 (25.8/87.5)	22.8 (85.1/13.2)	2.7 (21.6/1.4)	40.5	43.5
LLaMA-3-8B	3	0%	T	23.1 (63.9/14.1)	64.2 (70.8/58.7)	44.7 (40.1/50.5)	37.6 (23.5/94.7)	0.7 (21.2/0.3)	37.1	43.7
LLaMA-3-8B	0	25%	T	54.6 (54.9/54.4)	76.6 (74.4/78.9)	46.9 (62.4/37.6)	67.4 (65.9/69.1)	54.4 (50.8/58.5)	60.5	59.7
LLaMA-3-8B	0	50%	T	58.4 (60.3/56.7)	80.4 (78.3/82.7)	54.4 (65.5/46.5)	72.8 (70.5/75.3)	60.7 (57.2/64.6)	65.6	65.1
LLaMA-3-8B	0	75%	T	60.2 (61.5/59.0)	81.9 (78.6/85.4)	56.9 (68.2/48.9)	74.6 (73.8/75.4)	61.8 (58.7/65.1)	67.2	66.8
LLaMA-3-8B	0	100%	T	63.9 (66.1/61.9)	83.2 (81.4/85.1)	59.5 (66.0/54.1)	75.8 (76.7/74.9)	63.4 (59.8/67.5)	69.5	68.7
GPT-3.5	0	0%	T	43.6 (28.7/91.2)	29.2 (87.0/17.5)	29.4 (63.4/19.1)	26.3 (72.2/16.1)	39.2 (52.5/31.3)	37.9	35.1
GPT-3.5	0	0%	A	43.7 (28.8/90.9)	32.8 (87.3/20.2)	29.2 (61.1/19.2)	24.6 (75.7/14.7)	38.4 (51.1/30.7)	38.1	35.1
GPT-3.5	1	0%	T	45.9 (36.3/62.5)	63.3 (74.8/54.9)	49.8 (46.7/53.4)	48.8 (65.4/38.9)	38.0 (44.7/33.1)	48.3	48.6
GPT-3.5	3	0%	T	47.4 (43.1/52.7)	67.6 (69.8/65.6)	49.5 (40.4/63.9)	54.1 (45.4/66.8)	18.3 (41.4/11.7)	48.3	52.1
GPT-4	0	0%	T	43.1 (28.0/93.7)	28.1 (94.1/16.5)	27.4 (82.6/16.4)	29.4 (85.7/17.7)	37.8 (53.8/29.1)	37.5	34.7
GPT-4	0	0%	A	42.9 (27.7/95.3)	27.6 (94.3/16.1)	30.7 (79.2/19.0)	27.2 (79.5/16.4)	31.8 (51.5/23.0)	36.4	34.0
GPT-4	1	0%	T	51.1 (37.9/78.2)	58.9 (80.4/46.5)	55.3 (49.9/61.9)	54.5 (62.0/48.7)	27.0 (45.6/19.1)	49.0	50.9
GPT-4	3	0%	T	49.6 (39.8/65.8)	59.9 (81.7/47.2)	54.7 (53.5/56.1)	58.3 (52.5/65.4)	30.9 (40.5/24.9)	49.4	51.9

Table B2: F1(precision/recall), UA, and WA of LLMs on IEMOCAP under the 5-Way classification set-up. In table headers, “N” stands for the number of ICL samples in the prompt; “P” stands for the proportion of training data used for fine-tuning; “M” stands for the modality of input, either transcription (T) or ASR hypothesis (A).

Model	N	P	M	Neutral	Fearful	Dissatisfied	Apogetic	Abusive	Excited	Satisfied	MF1	WF1
GPT-2	0	0%	T	0.1 (100.0/0.0)	0.0 (0.0/0.0)	9.3 (5.6/27.8)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	2.8 (1.4/64.8)	31.4 (35.7/28.1)	7.3	24.0
GPT-2	1	0%	T	81.2 (69.8/97.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	23.3 (14.8/54.8)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	3.9	0.6
GPT-2	0	25%	T	82.4 (70.1/99.8)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	69.9 (71.4/68.5)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	11.7	1.9
GPT-2	0	50%	T	82.3 (70.0/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	69.6 (95.2/54.8)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	11.6	1.9
GPT-2	0	75%	T	82.4 (70.3/99.5)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	58.9 (47.9/76.7)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	9.8	1.6
GPT-2	0	100%	T	82.3 (70.0/99.8)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	68.1 (74.2/63.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	11.4	1.9
LLaMA-7B	0	0%	T	82.1 (69.7/100.0)	0.0 (0.0/0.0)	0.3 (33.3/0.2)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	6.3 (75.0/3.3)	0.0 (0.0/0.0)	1.1	0.3
LLaMA-7B	0	0%	A	82.1 (69.7/100.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0	0.0
LLaMA-7B	1	0%	T	83.0 (78.1/88.5)	26.1 (60.0/16.7)	2.6 (47.1/1.3)	0.0 (0.0/0.0)	57.9 (52.4/64.7)	16.0 (9.2/58.2)	59.0 (74.1/49.0)	26.9	42.6
LLaMA-7B	3	0%	T	27.9 (81.2/16.9)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	0.0 (0.0/0.0)	39.2 (24.4/99.2)	6.5	27.2
LLaMA-7B	0	25%	T	93.9 (91.5/96.4)	26.1 (60.0/16.7)	55.2 (81.6/41.7)	72.3 (93.5/58.9)	11.1 (100.0/5.9)	43.6 (69.0/31.9)	90.9 (89.1/92.7)	49.9	79.5
LLaMA-7B	0	50%	T	94.4 (93.1/95.8)	41.7 (83.3/27.8)	68.6 (80.0/60.1)	75.8 (92.2/64.4)	64.0 (100.0/47.1)	51.4 (69.8/40.7)	91.1 (89.8/92.5)	65.4	83.6
LLaMA-7B	0	75%	T	93.8 (93.0/94.5)	35.3 (37.5/33.3)	61.7 (84.9/48.5)	57.4 (41.5/93.2)	69.2 (100.0/52.9)	50.6 (54.4/47.3)	90.8 (88.7/93.1)	60.8	81.3
LLaMA-7B	0	100%	T	94.2 (93.3/95.2)	43.8 (50.0/38.9)	68.1 (78.3/60.3)	75.4 (93.9/63.0)	69.2 (100.0/52.9)	51.3 (63.9/42.9)	90.7 (88.8/92.6)	66.4	83.2
Alpaca-7B	0	0%	T	65.4 (85.3/53.1)	1.9 (1.1/11.1)	24.8 (28.5/22.0)	46.0 (85.2/31.5)	0.0 (0.0/0.0)	18.1 (13.8/26.4)	53.3 (38.7/85.7)	24.0	44.6
Alpaca-7B	0	0%	A	65.5 (83.0/54.1)	2.0 (1.1/11.1)	22.9 (26.2/20.4)	52.9 (93.1/37.0)	0.0 (0.0/0.0)	17.3 (15.4/19.8)	50.7 (37.2/80.0)	24.3	42.6
Alpaca-7B	1	0%	T	75.4 (81.4/70.2)	3.4 (1.9/22.2)	13.8 (23.8/9.8)	26.2 (100.0/15.1)	30.0 (100.0/17.6)	8.9 (4.8/65.9)	67.4 (69.4/65.6)	25.0	51.2
Alpaca-7B	3	0%	T	65.4 (85.3/53.1)	1.9 (1.1/11.1)	24.8 (28.5/22.0)	46.0 (85.2/31.5)	0.0 (0.0/0.0)	18.1 (13.8/26.4)	53.3 (38.7/85.7)	24.0	44.6
Alpaca-7B	0	25%	T	93.3 (91.9/94.7)	17.4 (40.0/11.1)	53.5 (72.5/42.4)	74.8 (92.0/63.0)	0.0 (0.0/0.0)	45.4 (64.0/35.2)	90.2 (86.6/94.2)	46.9	78.7
Alpaca-7B	0	50%	T	94.4 (93.1/95.8)	43.5 (100.0/27.8)	68.1 (79.9/59.3)	74.2 (83.1/67.1)	64.0 (100.0/47.1)	46.0 (66.7/35.2)	91.0 (89.6/92.6)	64.5	83.2
Alpaca-7B	0	75%	T	93.6 (91.0/96.4)	35.7 (50.0/27.8)	45.6 (90.6/30.5)	75.7 (79.1/72.6)	38.1 (100.0/23.5)	50.7 (67.3/40.7)	91.1 (88.7/93.6)	56.1	78.0
Alpaca-7B	0	100%	T	94.0 (92.0/96.1)	10.5 (100.0/5.6)	62.2 (76.4/52.5)	71.8 (73.9/69.9)	0.0 (0.0/0.0)	39.3 (77.4/26.4)	90.7 (90.2/91.1)	45.8	80.7
LLaMA-3-8B	0	0%	T	44.2 (79.5/30.6)	1.3 (0.6/55.6)	1.0 (13.0/0.5)	24.0 (14.6/67.1)	26.8 (15.8/88.2)	5.7 (2.9/81.3)	59.3 (59.6/59.1)	19.7	42.4
LLaMA-3-8B	0	0%	A	47.0 (80.7/33.1)	1.2 (0.6/50.0)	1.0 (12.5/0.5)	24.1 (14.7/67.1)	27.5 (16.3/88.2)	5.7 (3.0/78.0)	59.2 (59.3/59.1)	19.8	42.3
LLaMA-3-8B	1	0%	T	83.5 (76.3/92.3)	4.5 (3.8/5.6)	2.4 (16.0/1.3)	39.4 (30.8/54.8)	9.0 (4.7/100.0)	35.4 (46.4/28.6)	50.5 (87.4/35.5)	23.5	38.0
LLaMA-3-8B	3	0%	T	55.5 (85.1/41.1)	0.0 (0.0/0.0)	0.3 (20.0/0.2)	35.3 (62.1/24.7)	2.3 (1.2/100.0)	24.6 (60.9/15.4)	54.8 (39.2/90.9)	19.5	39.9
LLaMA-3-8B	0	25%	T	93.6 (90.7/96.7)	27.3 (75.0/16.7)	52.4 (87.6/37.4)	74.6 (97.8/60.3)	74.1 (100.0/58.8)	49.0 (64.3/39.6)	90.4 (89.5/91.3)	61.3	79.2
LLaMA-3-8B	0	50%	T	93.8 (91.1/96.7)	26.1 (60.0/16.7)	56.5 (86.1/42.1)	76.4 (94.0/64.4)	74.1 (100.0/58.8)	49.3 (62.7/40.7)	90.5 (90.1/91.0)	62.2	80.3
LLaMA-3-8B	0	75%	T	94.3 (92.4/96.3)	38.5 (62.5/27.8)	64.0 (85.2/51.3)	75.2 (83.3/68.5)	64.0 (100.0/47.1)	48.6 (66.0/38.5)	90.8 (89.5/92.0)	63.5	82.2
LLaMA-3-8B	0	100%	T	94.5 (92.4/96.7)	50.0 (100.0/33.3)	66.2 (85.2/54.1)	74.4 (85.7/65.8)	78.6 (100.0/64.7)	52.1 (69.1/41.8)	90.8 (90.5/91.2)	68.7	82.9
GPT-3.5	0	0%	T	82.8 (76.9/89.8)	20.7 (27.3/16.7)	8.2 (35.0/4.6)	61.9 (87.5/47.9)	61.5 (88.9/47.1)	31.6 (27.4/37.4)	50.0 (58.9/43.5)	39.0	40.0
GPT-3.5	0	0%	A	82.9 (76.8/90.0)	28.6 (100.0/16.7)	8.2 (35.4/4.6)	61.9 (87.5/47.9)	61.5 (88.9/47.1)	31.5 (27.2/37.4)	49.7 (58.8/43.1)	40.2	39.8
GPT-3.5	1	0%	T	66.0 (82.3/55.1)	36.4 (100.0/22.2)	13.5 (33.1/8.4)	7.9 (100.0/4.1)	75.7 (70.0/82.4)	10.7 (5.9/61.5)	56.0 (42.6/81.6)	33.3	43.2
GPT-3.5	3	0%	T	57.9 (82.1/44.7)	10.0 (50.0/5.6)	16.6 (34.4/10.9)	36.0 (100.0/21.9)	69.0 (83.3/58.8)	6.4 (3.4/71.4)	59.3 (46.5/81.8)	32.9	46.7
GPT-4	0	0%	T	88.3 (86.0/90.8)	50.0 (100.0/33.3)	16.4 (47.2/9.9)	52.5 (37.4/87.7)	74.1 (100.0/58.8)	42.2 (36.2/50.5)	79.0 (78.5/79.6)	52.4	62.3
GPT-4	0	0%	A	88.3 (82.7/94.6)	41.7 (83.3/27.8)	47.9 (70.1/36.4)	47.9 (33.2/86.3)	75.9 (91.7/64.7)	39.8 (31.6/53.8)	66.6 (89.4/53.1)	53.3	60.8
GPT-4	1	0%	T	78.8 (93.8/68.0)	41.7 (83.3/27.8)	52.5 (46.7/60.1)	42.7 (28.2/87.7)	83.3 (78.9/88.2)	14.6 (8.0/80.2)	71.8 (63.8/82.1)	51.1	64.4
GPT-4	3	0%	T	83.2 (91.9/76.0)	26.1 (60.0/16.7)	51.1 (48.0/54.6)	55.0 (42.0/79.5)	77.8 (73.7/82.4)	28.0 (20.3/45.1)	74.2 (63.0/90.4)	52.0	66.5

Table B3: F1(precision/recall), MF1 and WF1 of LLMs on EmoWOZ. In table headers, “N” stands for the number of ICL samples in the prompt; “P” stands for the proportion of training data used for fine-tuning; “M” stands for the modality of input, either transcription (T) or ASR hypothesis (A).

Model	N	P	M	Development Set		Test Set	
				Depressed	Not Depressed	Depressed	Not Depressed
GPT-2	0	0%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-7B	0	0%	T	47.5 (31.1/100.0)	11.4 (100.0/6.1)	52.2 (35.3/100.0)	8.3 (100.0/4.3)
LLaMA-7B	0	0%	A	45.9 (29.8/100.0)	0.0 (0.0/0.0)	51.1 (34.3/100.0)	0.0 (0.0/0.0)
LLaMA-7B	1	0%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-7B	3	0%	T	42.1 (27.9/85.7)	10.8 (50.0/6.1)	48.9 (33.3/91.7)	8.0 (50.0/4.3)
LLaMA-7B	0	25%	T	0.0 (0.0/0.0)	81.0 (69.6/97.0)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-7B	0	50%	T	0.0 (0.0/0.0)	81.0 (69.6/97.0)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-7B	0	75%	T	0.0 (0.0/0.0)	79.5 (68.9/93.9)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-7B	0	100%	T	0.0 (0.0/0.0)	76.3 (67.4/87.9)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
Alpaca-7B	0	0%	T	47.5 (31.1/100.0)	11.4 (100.0/6.1)	53.3 (36.4/100.0)	16.0 (100.0/8.7)
Alpaca-7B	0	0%	A	45.9 (29.8/100.0)	0.0 (0.0/0.0)	53.3 (36.4/100.0)	16.0 (100.0/8.7)
Alpaca-7B	1	0%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	15.4 (100.0/8.3)	80.7 (67.6/100.0)
Alpaca-7B	3	0%	T	45.9 (29.8/100.0)	0.0 (0.0/0.0)	51.1 (34.3/100.0)	0.0 (0.0/0.0)
Alpaca-7B	0	25%	T	12.5 (50.0/7.1)	82.1 (71.1/97.0)	0.0 (0.0/0.0)	77.2 (64.7/95.7)
Alpaca-7B	0	50%	T	11.8 (33.3/7.1)	80.5 (70.5/93.9)	0.0 (0.0/0.0)	77.2 (64.7/95.7)
Alpaca-7B	0	75%	T	10.5 (20.0/7.1)	77.3 (69.0/87.9)	13.3 (33.3/8.3)	76.4 (65.6/91.3)
Alpaca-7B	0	100%	T	18.2 (25.0/14.3)	75.0 (69.2/81.8)	0.0 (0.0/0.0)	75.0 (63.6/91.3)
LLaMA-3-8B	0	0%	T	47.1 (32.4/85.7)	37.2 (80.0/24.2)	43.2 (32.0/66.7)	36.4 (60.0/26.1)
LLaMA-3-8B	0	0%	A	46.4 (31.0/92.9)	21.1 (80.0/12.1)	42.9 (30.0/75.0)	14.3 (40.0/8.7)
LLaMA-3-8B	1	0%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-3-8B	3	0%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	0.0 (0.0/0.0)	79.3 (65.7/100.0)
LLaMA-3-8B	0	25%	T	27.0 (21.7/35.7)	52.6 (62.5/45.5)	29.6 (26.7/33.3)	55.8 (60.0/52.2)
LLaMA-3-8B	0	50%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	0.0 (0.0/0.0)	77.2 (64.7/95.7)
LLaMA-3-8B	0	75%	T	0.0 (0.0/0.0)	82.5 (70.2/100.0)	47.1 (36.4/66.7)	50.0 (69.2/39.1)
LLaMA-3-8B	0	100%	T	20.0 (33.3/14.3)	78.4 (70.7/87.9)	47.1 (36.4/66.7)	50.0 (69.2/39.1)
GPT-3.5	0	0%	T	54.5 (60.0/50.0)	79.2 (76.0/82.6)	64.3 (64.3/64.3)	84.8 (84.8/84.8)
GPT-3.5	0	0%	A	37.5 (75.0/25.0)	81.5 (71.0/95.7)	56.0 (63.6/50.0)	84.1 (80.6/87.9)
GPT-3.5	1	0%	T	13.3 (33.3/8.3)	76.4 (65.6/91.3)	40.0 (45.5/35.7)	78.3 (75.0/81.8)
GPT-3.5	3	0%	T	37.5 (75.0/25.0)	81.5 (71.0/95.7)	56.0 (63.6/50.0)	84.1 (80.6/87.9)
GPT-4	0	0%	T	63.6 (70.0/58.3)	83.3 (80.0/87.0)	59.3 (61.5/57.1)	83.6 (82.4/84.8)
GPT-4	0	0%	A	44.4 (66.7/33.3)	80.8 (72.4/91.3)	41.7 (50.0/35.7)	80.0 (75.7/84.8)
GPT-4	1	0%	T	80.0 (76.9/83.3)	88.9 (90.9/87.0)	55.6 (45.5/71.4)	72.4 (84.0/63.6)
GPT-4	3	0%	T	74.1 (66.7/83.3)	83.7 (90.0/78.3)	58.5 (44.4/85.7)	58.5 (44.4/85.7)

Table B4: F1(precision/recall) of LLMs on DAIC-WOZ. In table headers, “N” stands for the number of ICL samples in the prompt; “P” stands for the proportion of training data used for fine-tuning; “M” stands for the modality of input, either transcription (T) or ASR hypothesis (A).

## Chapter 8

# Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation

This chapter summarises our work on enhancing task-oriented dialogue systems with emotional behaviours across understanding, decision-making, and response generation, and gives a verbatim copy of our manuscript (Feng et al., 2024a):

Shutong Feng et al. (Sept. 2024a). “Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation”. In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara et al. Kyoto, Japan: Association for Computational Linguistics, pp. 699–717. DOI: 10.18653/v1/2024.sigdialog-1.60. URL: <https://aclanthology.org/2024.sigdialog-1.60/>

## 8.1 Summary

In previous Chapters 5, 6, and 7, we focus primarily on emotion recognition learning. In the work presented in this chapter, we step further to integrate emotion modelling throughout the full dialogue pipeline. Specifically, we give our emotional task-oriented dialogue systems the ability to understand user emotions, to generate emotionally appropriate responses, and to regulate system behaviours.

We extended the EmoWOZ dataset with affective behaviour labels for system utterances, resulting in EmoWOZ 2.0. This expanded resource enables the modelling of the system’s affective behaviour, which gives the system the ability to respond emotionally. We then set up systems with distinct architectures: (1) a modular dialogue system with a policy optimised jointly for task success and emotional appropriateness via reinforcement learning, and (2) an end-to-end large language model-based system trained to generate responses conditioned on both task information and emotional context via supervised learning. To support evaluation, we also utilise an emotion-aware user simulator capable of producing both user utterances and emotional reactions in natural language.

We evaluate our systems using the corpus, an emotion-aware user simulator, and real humans. Our results show that emotion-integrated systems significantly improve task success, user satisfaction, and hallucination compared to emotion-agnostic baselines in both modular and end-to-end systems. Our findings highlight the effectiveness of closing the emotional loop in dialogue and affirm the importance of emotion in task-oriented dialogues.

## 8.2 Personal Contributions

I contributed to the construction of the EmoWOZ 2.0 dataset, the incorporation of emotion recognition to dialogue state tracking, the development of the emotional natural language generator, the development, training and evaluation of end-to-end models, the hallucination evaluation, the setup of human trials, and paper writing. Hsien-Chin Lin contributed to the development of the user simulator, the

training of the modular dialogue policy, and the corpus-based and simulator-based evaluation of modular systems. Christian Geishauser contributed to the development and training of the modular dialogue policy. Milica Gasić supervised the project. Co-authors assisted in writing and proofreading.

# Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation

Shutong Feng\*, Hsien-chin Lin\*, Christian Geishauser\*, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, Milica Gašić

Heinrich Heine University Düsseldorf, Germany

{fengs, linh, geishaus, lubis, niekerk, heckmi, ruppik, revuk100, gasic}@hhu.de

## Abstract

Emotions are indispensable in human communication, but are often overlooked in task-oriented dialogue (ToD) modelling, where the task success is the primary focus. While existing works have explored user emotions or similar concepts in some ToD tasks, none has so far included emotion modelling into a fully-fledged ToD system nor conducted interaction with human or simulated users. In this work, we incorporate emotion into the complete ToD processing loop, involving understanding, management, and generation. To this end, we extend the EmoWOZ dataset (Feng et al., 2022) with system affective behaviour labels. Through interactive experimentation involving both simulated and human users, we demonstrate that our proposed framework significantly enhances the user’s emotional experience as well as the task success.

## 1 Introduction

In recent years, conversational artificial intelligence (AI) has become increasingly prevalent in various domains, providing users with interactive and personalised experiences. Emotions play a crucial role in human communication and can influence the way individuals perceive, process, and react to information (Ekman, 1992). Consequently, incorporating emotions into conversational AI has emerged as a promising avenue for improving user experience and creating more human-like interactions (Picard, 2000).

Task-oriented dialogue (ToD) systems, an important genre of conversational AI, are designed to assist users in fulfilling specific tasks or queries. In contrast to chit-chat or open-domain dialogue systems, which focus on creating engaging and entertaining conversations, ToD systems interact with users in a more structured way with a clear objective under specific domains (Jurafsky and Martin,

2009). While significant advancements have been made in natural language processing and ToD systems, there remains a critical challenge in creating systems that can understand and respond to not only the informational needs of users but also their emotional states.

In ToD, emotion is centred around the user goal, making it more contextual and subtle (Feng et al., 2022). A recent study has shown that the valence of user emotion in ToD positively correlates with dialogue success (Lin et al., 2023). This observation aligns with a number of emotional theories. For example, the appraisal theory of emotion argues that emotion is the result of our evaluation of a situation (Arnold, 1960; Lazarus, 1966). In relation to a ToD user goal, it is straightforward to see how task fulfilment would lead to positive emotions and failures to negative ones. Similarly, the Ortony-Clore-Collins (OCC) model of emotion states that emotion is the result of elicitation by events, agents, and objects (Ortony et al., 1988). Feng et al. (2022) have drawn the connection between the OCC model and user emotions in ToD. Therefore, besides inferring emotional states from dialogue utterances, an agent also needs to reason about emotion-generating situations and to utilise this information to achieve dialogue success.

The integration of emotion into the full ToD pipeline has been a long-standing interest (Bui et al., 2010; Ren et al., 2015). Yet, early works explored analytical solutions in constrained set-ups, which hindered their applications in more complicated scenarios. Recently, a number of resources emerged for studying user affect in ToDs, e.g. emotion, sentiment, or satisfaction (Mendonca et al., 2023; Feng et al., 2022). This has motivated efforts to model user emotion via data-driven approaches, such as emotional user simulation (Lin et al., 2023) and user emotion recognition (Feng et al., 2023a; Stricker and Paroubek, 2024). However, to the best of our knowledge, no work so far has com-

\*These authors contributed equally to this work.

bined these emotional aspects into a fully-fledged dialogue system and an interactive pipeline where emotions play a role in understanding, generation, as well as management of the conversation.

To achieve this, we need to endow the dialogue system with the ability to respond with an affective behaviour, closing the emotional loop between the user and the system in ToDs. Towards this goal, we make the following contributions:

- We extend EmoWOZ, a large-scale ToD dataset for user emotions (Feng et al., 2022), with annotations for *affective conduct* in 71k system utterances. To the best of our knowledge, this is the first large-scale and open-source corpus dedicated to the system’s affective behaviour in ToDs.
- We incorporate emotion in the complete ToD interaction loop for understanding, management, and generation by building a modular system around an *emotion-aware* and *emotion-expressive* policy. We also build an emotional LLM-based end-to-end ToD system that involves emotion in understanding and generation.
- For our modular system, we train our dialogue policy via reinforcement learning (RL) on the natural language level, leveraging emotions and task success as reward signals. We train the end-to-end system on our newly collected dataset via supervised learning (SL). For both systems, we show through interactive evaluation that emotion in the ToD loop can enhance user’s emotional experience as well as the task success. This highlights the importance of modelling emotions in ToDs.

## 2 Related Work

In this section, we discuss related works on incorporating emotion in each stage of ToD pipeline: understanding, management, and generation. These stages are modelled explicitly with multiple models in modular systems and implicitly with a unified model in end-to-end systems (Hosseini-Asl et al., 2020; Stricker and Paroubek, 2024).

### 2.1 Understanding User Emotion

Modular ToD systems rely on natural language understanding (NLU) and dialogue state tracking (DST) modules to translate and accumulate semantic concepts related to user goals. Typically, these

semantic concepts are strictly limited to those defined in the ontology, i.e. domains, slots, and values the system can talk about.

Given its potential as an important piece of information for the system’s subsequent decision-making, emotion can be considered as part of the dialogue state. Feng et al. (2022) showed that multi-task training a DST model for emotion recognition simultaneously improves its joint goal accuracy, suggesting the complementarity between DST and emotion recognition in conversation (ERC). Recently, Stricker and Paroubek (2024) modelled user emotion as an intermediate task in end-to-end ToD systems and improved overall system performance. Standalone ERC models dedicated to ToDs (Li et al., 2023; Feng et al., 2023b) can be used in modular systems in parallel with any DST to extend the dialogue state with user emotions.

### 2.2 Dialogue Management with Emotion Feedback

In ToD, one way to train the dialogue policy is via RL to maximise task success, indicated at the end of the dialogue based on user goal fulfilment (Levin and Pieraccini, 1997; Kwan et al., 2023). Since user emotion is highly associated with task success (Lin et al., 2023), it is intuitive to leverage user emotion during the dialogue for providing more dense and diverse reward signals. Bui et al. (2010) incorporate user emotion into the policy state by modelling affective dialogue management through a factored partially-observable Markov decision process (POMDP) and analytically find an optimal policy. This is however neither feasible for larger problems, nor has this been integrated in interactive set-ups. Zhang et al. (2021) addressed the delayed reward problem in dialogue policy learning with a predefined emotion-based turn-level reward. Zhu et al. (2024) consider the difference between the user’s positive emotion intensity and the next turn’s emotion utility value for top-k action selection. We take a step further by incorporating emotion in policy state *and* reward function. We then leverage emotion in RL to find optimal semantic actions *and* affective expression of the system, which has not been explored before.

### 2.3 Generating Affective Response

The natural language generation (NLG) module in ToD systems realises semantic actions from the policy into natural language. Traditionally, ToD NLG focuses on translating task-related semantic

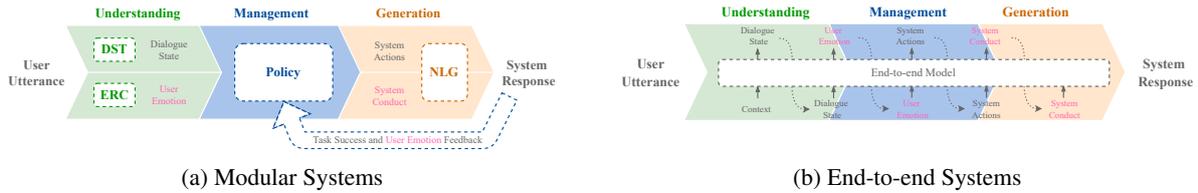


Figure 1: Infusing emotions into modular and end-to-end ToD systems.

actions and overlooks other aspects of system responses such as emotion. There have been efforts to create datasets which help enrich ToD system responses with chit-chat (Sun et al., 2021; Chen et al., 2022; Young et al., 2021). Stricker and Paroubek (2024) attempted to refine end-to-end system output with a large language model (LLM) under a chain-of-thought framework to enhance simulated system empathy. Different from previous works, we aim to enrich system response with the subtle affective conduct jointly with dialogue actions in a fully controllable approach.

## 2.4 Simulating User Emotional Behaviour

User simulators (USs) simulate user behaviour in ToDs. Although they are not a part of the system, they play essential roles in training dialogue policy via RL and serving as an efficient evaluation platform for dialogue policy (Eckert et al., 1997). Most existing USs focus on modelling user’s behaviour in terms of semantic actions and natural language by taking system semantic actions (Kreyszig et al., 2018; Lin et al., 2021, 2022). Zhang et al. (2021) built a US that additionally incorporated handcrafted emotion transitions in different situations. Kim and Lipani (2022) used a data-driven approach and simulated satisfaction levels along with the intent and the utterance. Lin et al. (2023) further proposed data-driven EmoUS to model more nuanced user emotions with enhanced controllability via user persona settings. This motivates us to move one step further to capture more fine-grained affective expressions of the system from natural language response directly.

## 3 EmoWOZ 2.0: A Fully Emotion-annotated ToD Dataset

To study emotion in real-world interactions between users and human operators in the ToD setting, we extend EmoWOZ (Feng et al., 2022) by further annotating the *affective behaviour* of the system, which is acted by human-beings. We call

this dataset with extended labels *EmoWOZ 2.0*. \*

In ToDs, the user and the system play different roles. Users may express a wide range of emotions during interactions based on their goals and experiences with the operator. The system is responsible for managing and facilitating the conversation and is supposed to behave professionally and politely to achieve the goal. Therefore, it is necessary to consider different sets of affective behaviours in the user and the system respectively. We refer to the concept of the operator’s affective behaviour as **affective conduct**, or **conduct** for short.

**Annotation Scheme** According to studies on customer satisfaction in business (Grandey et al., 2011), competent operators in ToD try to guide user emotion towards positive valence by making use of subtle emotion in their response while providing correct information. By considering the set of user emotions in EmoWOZ and the OCC emotion model (detailed justification in Appendix A.1), we arrive at five affective conduct classes:

- **Neutral:** the operator does not explicitly make use of any affective conduct.
- **Compassionate:** the operator is sympathetic about user’s situation, usually in response to a fearful/disappointed user in an unpleasant situation.
- **Apologetic:** the operator apologises for their mistake, usually in response to a dissatisfied user.
- **Enthusiastic:** the operator is feeling happy for the user or showing extra eagerness to help. This conduct takes place usually in response to a neutral or excited user.
- **Appreciative:** the operator acknowledges the – at least partial – task success, usually signalled as user’s satisfaction.

**Annotation Set-up** We annotated the conduct for all operator utterances in the MultiWOZ subset of

\*EmoWOZ 2.0 is released under CC By 4.0 NC license, following the original EmoWOZ release. The dataset can be found at <https://gitlab.cs.uni-duesseldorf.de/general/dsml/emowoz-2.0-public/>

EmoWOZ. Machine-generated system responses in the DialMAGE subset came from a template NLG, which we considered to have neutral conduct because those templates aimed to express actions concisely rather than conveying emotions by design.

We followed the data collection and quality assurance set-up of EmoWOZ and conducted the annotation via the Amazon Mechanical Turk platform. Details and an illustration of the annotation interface can be found in Appendix A.2.

**Annotation Quality** Each utterance has been annotated by at least three annotators. The inter-annotator agreement as measured with Fleiss’ Kappa is 0.647, suggesting substantial inter-annotator agreement. The annotator confusion matrix and label distribution can be found in Appendix A.3 and A.4, respectively.

## 4 Infusing Emotions into ToD Systems

We propose to incorporate emotion into the full interactive ToD pipeline, which is primarily comprised of three stages: understanding, management, and generation. We aim for understanding to accurately recognise the user’s emotion in addition to the task-centred dialogue state. For dialogue management, we make use of emotion for optimal action selection. Lastly, we additionally condition the natural language generation on the system conduct to generate more diverse and emotion-aware responses. These are realised in each modular system component individually (Section 4.1 to 4.3) and as intermediate tasks in the unified model in end-to-end systems (Section 4.4).\*

### 4.1 Expanding Dialogue State with Emotion

In our modular system, we use an ERC model in parallel with a DST model. This allows a flexible selection of DST and the associated ontology. The inferred user emotion is appended to the dialogue state.

For ERC, we use the ContextBERT-ERToD model (Feng et al., 2023a) as our user emotion recognition front-end because of its good ERC ability and fast inference. It is a BERT-based classification model (Devlin et al., 2019) that considers dialogue context and state in addition to the user

---

\*The code of pipeline systems, end-to-end systems, and the user simulator can be found at <https://gitlab.cs.uni-duesseldorf.de/general/dsml/emoloop-public/>

utterance. It reports a weighted F1 score of 83.9% for emotions excluding neutral.

For DST, we use the SetSUMBT model (van Niekerk et al., 2021). This model, based on the RoBERTa language model (Liu et al., 2019) and a recurrent context tracker adopts a picklist approach to DST. Specifically, we employ the Ensemble-Distribution-Distilled variant of SetSUMBT, a refined version that distils knowledge from an ensemble of models. This version reports a joint goal accuracy of 51.22% on MultiWOZ. The architectural design of SetSUMBT also allows transferability to new domains, and such an ability has been exemplified with a transformer-based dialogue policy under a continual learning set-up (Geishauser et al., 2024).

### 4.2 Emotion-aware Dialogue Policy

For dialogue management in the modular system, we build a dialogue policy that considers the user emotion in the input and produces an emotion-augmented system output. We utilise the dynamic dialogue policy transformer (DDPT) architecture (Geishauser et al., 2022) since it was built for optimising dialogue policies that require extendable input and output, which facilitate the adaptation to new domains and ontologies. The dialogue policy leverages emotions in three ways: considering user emotion in the input, generating system affective conduct in the output, and considering user emotion in the reward for RL.

**Emotion Input and Output** The user emotion, as a part of the dialogue state, is incorporated into the dialogue state through embedding the perceived user emotion with RoBERTa. For semantic action selection, DDPT produces a sequence of domain-intent-slot triplets auto-regressively through its transformer decoder, e.g. `restaurant-inform-phone`, `restaurant-request-food`, until a stop token is generated. In order to predict *emotional* system conduct, after DDPT outputs the semantic actions, we decode the sequence for one more step to generate the system conduct action, considering the perceived user emotion from the dialogue state.

**Emotion Augmented Reward** We incorporate user emotion into the reward for RL by considering the associated sentiment. More specifically, we define  $c(\text{satisfied}) = 1$ ,  $c(\text{dissatisfied}) = c(\text{abusive}) = -1$ ,  $c(\text{neutral}) = 0$ . For the remaining user emotions that are not elicited by the

system, we set  $c(\text{emotion}) = 0$ . For any emotion  $e$ , we multiply  $c(e)$  by a hyperparameter  $\beta$  to weight the influence of emotion in the reward. Note that utilizing  $\beta \cdot c(e)$  directly could encourage the dialogue policy to produce long dialogues with unnecessary turns as long as they produce positive user sentiment. In order to prevent this, we shift  $\beta \cdot c(e)$  such that it is at most 0 by defining the emotion reward for an emotion  $e$  as  $r_{\text{emo}}(e) = \beta \cdot c(e) - \beta$ .

The emotional reward is combined with the standard reward  $r_{\text{task}}$  in dialogue policy learning that equals  $-1$  in every non-terminating turn for encouraging efficiency and either  $-T$  or  $2T$  for dialogue failure or success, where  $T$  denotes the maximum permitted number of turns. The final reward is thus given by  $r = r_{\text{task}} + r_{\text{emo}}$ . We refer to this policy with expanded dialogue state input, expanded dialogue action output, and emotion reward as **EmoD-DPT**.

### 4.3 Expressing Emotion in Response

Our modular system NLG was built based on the BART model (Lewis et al., 2020). We followed existing works to formulate the ToD NLG problem as a sequence-to-sequence task (Peng et al., 2020; Zhu et al., 2023) where the input is a sequence containing semantic concepts in textual form (e.g. tuples of [intent, domain, slot, value]), and the output is natural language conveying the semantic meaning. Our model input consists of the user utterance, system semantic actions, and the system conduct. We refer to our system NLG as **SEC-BART**: a both semantically and emotionally conditioned BART. In our ablation study, we used **SC-BART**, the version that is only conditioned on the semantic actions in the non-emotional ToD pipeline.

On MultiWOZ, SEC-BART achieves a BLEU score of 34.9 and a slot error rate of 3.6%, comparable to existing SOTAs (Peng et al., 2020; Zhu et al., 2023). Details of model training and performance can be found in Appendix C.

### 4.4 Emotional End-to-end System

We follow the work of Stricker and Paroubek (2024), where ERC is added as an intermediate task in the end-to-end ToD modelling, i.e. emotion is incorporated in the understanding stage. We further consider emotion in the generation stage by predicting the system conduct in the end-to-end pipeline, as illustrated in Figure 1b. To this end, we build a LLaMA-based end-to-end ToD system that involves emotion in both understanding

and generation, with LLaMA-2-7B (Touvron et al., 2023) as the backbone. As illustrated in Figure 1b, it takes dialogue history and the recognised user emotion as input, and then auto-regressively generates the dialogue state, user emotion, system actions, system conduct, and delexicalised natural language response. The response is then lexicalised via database queries based on the intermediately generated dialogue state and system actions. We refer to this end-to-end model as **EmoLLAMA**.

We did not train EmoLLAMA via RL with task and emotion feedback from the user simulator because it would take more than 20 days on an A100 40GB to simulate the same number of dialogues as we did to train the EmoDDPT policy in the modular system. We therefore leave efficient training of LLM-based ToD systems via RL as a future research direction.

### 4.5 Emotional User Simulation

Traditionally, user simulators interact with the system on the semantic level for efficiency. To capture more fine-grained expressions of system conducts in natural language, we build **langEmoUS** based on EmoUS (Lin et al., 2023). langEmoUS interacts with the system on the natural language level, e.g. it takes the system utterance, user goal, turn information and user persona as inputs and generates user emotion and user utterance. The turn information represents the dialogue progress, i.e. the turn number. Following the setting in Lin et al. (2023), the user persona is extracted from the dialogue history, e.g. if a user is excited to visit a museum in the conversation, then its persona is  $\{\textit{attraction} : \textit{excited}\}$ , when training the user model supervisedly. During inference, the user persona is sampled from the distribution of the corpus.

LangEmoUS achieves macro F1 scores of 0.742 and 0.521 for user sentiment prediction and emotion prediction, respectively, significantly outperforming existing state-of-the-art models (Kim and Lipani, 2022; Lin et al., 2023) (see Appendix B).

## 5 Experimental Set-up

### 5.1 Modular System Set-up

**EmoLoop** This is our proposed modular system with emotion incorporated for understanding, management, and understanding, as outlined in Figure 1a and Figure 2. It includes the following modules: SetSUMBT DST, ContextBERT-ERToD ERC, EmoDDPT policy, and SEC-BART NLG.

EmoDDPT is trained via RL on the natural language level with langEmoUS.

**SimpleLoop** This is the non-emotion baseline to EmoLoop. It neither predicts user emotion for the state, uses emotion reward to train the policy, nor generates system conduct for emotional response generation. Specifically, it includes the following modules: SetSUMBT DST, DDPT policy, and SC-BART NLG. DDPT is trained via RL on the natural language level with langEmoUS.

### 5.1.1 Dialogue Policy Optimisation

We implement our system in the ConvLab-3 framework (Zhu et al., 2023). We pre-trained the policy on MultiWOZ 2.1 (Eric et al., 2020), followed by online RL through interaction with our US. During RL, in addition to the emotion reward as outlined in Section 4.2, we set the task reward as  $-1$  in every turn to encourage efficiency, and  $80$  or  $-40$  for dialogue success or failure. A dialogue is successful if the system provides the requested information to the user and books the correct entities (if possible). For emotional reward, we set  $\beta = 2$ . We pre-train each policy on MultiWOZ, followed by 15k dialogues with langEmoUS via RL for 6 random seeds. For every 1k dialogues of training, we evaluate the policy for 500 dialogues. We use overall return to select the best checkpoint. All peripheral modules were trained, implemented, and evaluated in the ConvLab-3 environment.

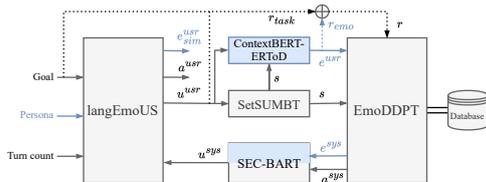


Figure 2: RL training set-up for EmoDDPT.

**Language-level RL Training** As illustrated in Figure 2, our policy, EmoDDPT, interacts with langEmoUS on the natural-language level where the policy actions and conduct ( $a^{sys}$ ,  $e^{sys}$ ) is realised into natural language,  $u^{sys}$  with SEC-BART. The US takes natural-language input and outputs natural-language user utterances  $u^{usr}$  after auto-regressively generating the simulation target user emotion  $u_{sim}^{usr}$  and user actions  $a^{usr}$ . The perceived user emotion  $e^{usr}$  and dialogue state  $s$  are determined by ContextBERT-ERToD and SetSUMBT respectively.

## 5.2 End-to-end System Set-up

**EmoLLAMA** This our proposed end-to-end system as described in Section 4.4.

**SimpleLLAMA** This is the non-emotional baseline, which is also used in the work of Stricker and Paroubek (2024). Compared with EmoLLAMA, it does not consider user emotions as a part of the model input, nor does it auto-regressively predict user emotion and system conduct.

Both EmoLLAMA and SimpleLLAMA are trained and evaluated with EmoWOZ 2.0 using the environment provided by Stricker and Paroubek (2024) and following default parameters. Their interactive evaluations were set up in the ConvLab-3 environment.

## 5.3 Evaluation

**Corpus Evaluation** We report *inform* and *success* rates. Inform rate evaluates if the system provides entities from the database that fulfill user’s constraints. Success rate assesses if the system delivers all information requested by the user. To generate each system response, the ground-truth dialogue history was used as system input.

**Interactive Evaluation** For interactive evaluation, our systems interact with langEmoUS. We report the *success* rate and the average user *sentiment* in simulated dialogues to account for user emotional experience. Specifically, the turn-level sentiment score is  $+1$  if the user emotion is positive,  $0$  if neutral and  $-1$  if negative. User sentiment is determined by the ERC.

**Human Trial** We set up a human trial using the DialCrowd toolkit (Huynh et al., 2022) on the Amazon Mechanical Turk platform. We set up two pairs of comparison: 1) SimpleLLAMA vs. EmoLLAMA and 2) SimpleLoop vs. EmoLoop. Volunteers are presented with randomly generated single or multi-domain goals. A goal contains a set of constraints for entities that the user should be looking for (e.g. the price range and the location of a restaurant) and specifies the information they should extract from the system (e.g. the phone number and booking reference of the restaurant). Given a goal, volunteers would need to talk to each system to fulfill the goal. They then give ratings to each of them based on objective (whether the goal has been fulfilled) and subjective metrics (how they feel about the system). Survey questions include objective task success and subjective user

System	Type	Corpus		User Simulator		Human	
		Inform	Success	Success	Sentiment	Success	Sentiment Rating
SimpleLLAMA	End-to-end	0.785	0.705	0.330	0.214	0.819	3.97
EmoLLAMA	End-to-end	<b>0.833</b>	<b>0.760</b>	0.342	<b>0.250</b>	<b>0.894</b>	<b>4.16</b>
SimpleLoop	Modular	0.700	0.621	0.556	0.337	0.798	3.85
EmoLoop	Modular	<b>0.753</b>	0.635	0.531	<b>0.405</b>	<b>0.917</b>	<b>4.15</b>

Table 1: System evaluation, including corpus-based evaluation, interaction with user simulator and human trial. Values in bold mean best scores with statistically significant difference  $p < 0.05$ .

sentiment. Details of the website interface and survey questions can be found in Appendix D. To obtain more reliable ratings, we filtered out dialogues with poor quality, e.g. containing very short user utterances or non-natural language, and with inconsistent ratings, e.g. system A had better rating in all aspects but overall the rater found system B better. Overall, we collected 203 valid ratings for the SimpleLLAMA-EmoLLAMA comparison and 253 for the SimpleLoop-EmoLoop comparison from 40 unique raters.

## 6 Results and Discussion

### 6.1 Corpus Evaluation

Although it is not a common practice to evaluate RL-trained modular ToD systems on a corpus, we provide such results for a basic understanding and comparison with end-to-end systems. Our goal is not beating SOTA on task-related metrics, but examining interactive abilities of the system and the role of emotion in it. As shown in Table 1, incorporating emotion significantly improves inform rate of both types of systems and success rate of the end-to-end system.

It is not surprising that modular systems underperform when compared with end-to-end systems. Modular systems are trained via RL, which allows the policy to explore more diverse dialogue trajectories but diverges from what a policy can learn from the corpus only. This reflects the limitation of corpus evaluation in accounting for ToD system performance, as pointed out by Lubis et al. (2022).

### 6.2 Evaluation with User Simulator

In interactive evaluation, both EmoLoop and EmoLLAMA perform significantly better in terms of average sentiment than their respective non-emotional baseline while maintaining the same level of success rate. For end-to-end models, despite the fact that they are not optimised via RL with the simulated user, the average sentiment in the simulated user also improves significantly.

When comparing performance across system types, modular systems perform better than end-to-end models on task success and simulated user sentiment since modular system policies have been optimised for the simulated user via RL. SimpleLLAMA and EmoLLAMA, trained via SL only, cannot adequately cope with the more diverse user goals and situations of the simulated user. This motivates our future work to leverage the simulated user and to train end-to-end systems via RL.

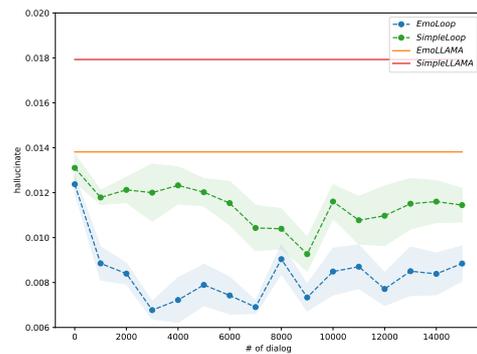


Figure 3: The average hallucination rate of modular systems during RL training with langEmoUS. For end-to-end systems, we report hallucination rate after SL.

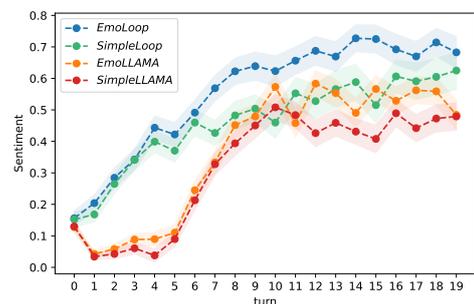


Figure 4: Average sentiment at different turn positions during language-level interaction with langEmoUS.

**Hallucination** In ToD, a hallucination is defined as a value in the system response that is not supposed to be informed according to system actions. As shown in Figure 3, the hallucination rate of each

type of systems is improved as emotion is incorporated into the pipeline. The hallucination rate is lowered from 1.8% for SimpleLLAMA to 1.4% for EmoLLAMA. We observe that end-to-end systems are more prone to the hallucination problem than modular systems as slot placeholders in the delexicalised end-to-end system response do not always match the intermediately generated dialogue actions. Hallucination rates of SimpleLoop and EmoLoop are around 1.3% at the beginning of the interactive RL training and continue to improve as the RL progresses.

### Progression of User Sentiment in Dialogues

Figure 4 shows the average sentiment of langEmoUS at each turn of interactions with our systems. The sentiment level of langEmoUS becomes more positive as the dialogue progresses and moves towards user goal completion in all systems. The primary difference between modular systems and end-to-end systems is that in earlier turns, modular systems are able to satisfy the simulated user better, as illustrated in higher and more positive sentiment level before turn 8.

### 6.3 Human Trials

We carried out human trials to compare two pairs of systems in Table 1. Within each pair of comparison, the emotion-incorporating model significantly outperforms its non-emotion version in terms of both the success rate and user sentiment. This further confirms our findings from corpus and user simulator evaluations. Example dialogue excerpts are given in Appendix D.3 to exemplify how emotional ToD systems made use of affective conduct in case of neutral and unsuccessful interactions.

Although human ratings across system types are not directly comparable, it is noteworthy that the absolute improvement from SimpleLLAMA to EmoLLAMA ( $\Delta\text{Success} = 0.075$ ,  $\Delta\text{Sentiment} = 0.19$ ) is smaller than that from SimpleLoop to EmoLoop ( $\Delta\text{Success} = 0.119$ ,  $\Delta\text{Sentiment} = 0.30$ ). Such difference can be attributed to the lack of RL training in LLM-based systems.

### 6.4 Ablation Study

We ablate our emotional modular and end-to-end systems by incorporating emotion in different parts of the pipeline. Table 2 summarises their interactive performance with langEmoUS.

For both modular systems and end-to-end systems, incorporating emotion does not significantly

System	Und	Gen	Man	Success	Sentiment
SimpleLLAMA	-	-	-	0.330	0.214
	+	-	-	0.360	0.233
EmoLLAMA	-	+	-	0.373	0.229
	+	+	-	0.342	<b>0.250</b>
SimpleLoop	-	-	S	0.556	0.337
	+	-	S+E	0.559	0.354
EmoLoop	-	+	S	0.543	0.361
	+	+	S+E	0.531	<b>0.405</b>

Table 2: Success and average user sentiment of systems from the interactive evaluation with langEmoUS. +/- means whether emotion is involved in the corresponding ToD stage: **U**nderstanding, **M**anagement, or **G**eneration. For Management, “-” means the system is trained via SL, “S” and “E” mean training via RL with success reward and emotion reward respectively.

change task success with the user simulator ( $p > 0.5$ ). The average user sentiment does improve slightly as emotion is introduced in understanding (plus management) and generation. Yet, the improvement from the non-emotional base system only becomes significant when emotion is added to all ToD stages. This highlights the importance of considering emotion in the whole ToD loop: it is necessary not only to understand user emotion but also to make use of it for dialogue management and respond with the appropriate conduct.\*

Figure 5 illustrates the change in the average sentiment of the simulated user during RL. At the beginning, average sentiments of modular systems fall in the similar range as SL-trained end-to-end systems, and are then further improved by RL. This highlights the importance of task success and emotion feedback signal for RL in ToD systems.

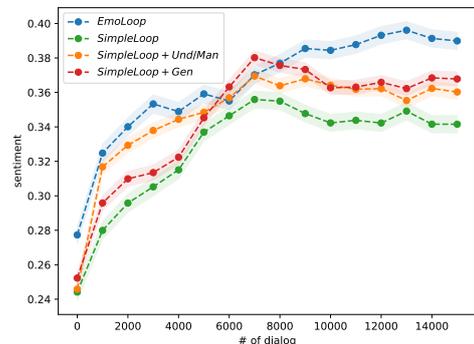


Figure 5: The average sentiment of langEmoUS during RL training of modular policy.

\*See Appendix E.1 for ablation study on EmoLoop with SL policy. A similar trend has been observed.

## 7 Conclusion

In this work, we incorporate emotion into the complete ToD processing loop, involving understanding, management, and generation. To achieve this, we first enrich the EmoWOZ dataset with system conduct labels to construct EmoWOZ 2.0. We then build modular and end-to-end ToD systems, as well as emotional user simulators with the newly collected dataset. We train the modular system policy via RL with the emotional user simulator and the end-to-end system via SL on EmoWOZ 2.0. Through interactive evaluation with both simulated and human users, we show that incorporating emotion into ToD systems can improve user’s emotional experience as well as task success.

There is still a long way to go from our work to the perfect emotional ToD system. Yet, we show our method as a promising avenue to achieve this ultimate goal. In our study, we directly translate user emotion labels into valence scores on a linear scale as a reward for RL. We believe that utilising the full set of user emotion labels for diverse reward would be a promising future direction.

We hope that with our work, we can motivate future research efforts to look at user experience beyond task success for ToDs and bring about insights to other task-oriented conversation settings. We would also like to highlight the opportunities in further improving LLM-based end-to-end ToD systems via RL, combining established approaches for policy training in modular systems and recent advancements in LLM research in other applications.

## 8 Limitations

One of the main limitations of modular ToD systems is the error accumulation in the pipeline for both modular and end-to-end systems. In modular systems, since each module is trained with a dataset associated with a limited ontology, the concepts that the system can understand and express are also limited. Although the DDPT policy, SetSUMBT DST, and many other models such as Trippy-R (Heck et al., 2022) are built with the ability to handle out-of-domain requests, the generalisability and robustness of ToD systems are still challenges in the field that is yet to be solved.

All system modules have been trained in a supervised fashion on EmoWOZ 2.0. Therefore, the dataset contains limited dialogue situations and inherent bias. As seen in the dialogue examples in

the appendix, the emotional responses are also limited. Yet, EmoWOZ 2.0 is the best resource we have at the moment. Data augmentation has been applied when training the NLG and the ERC model to mitigate the lack of diversity in the dataset. The RL training of the policy also allows the policy to explore more diverse dialogue trajectories. For the user simulator, considering data augmentation and more attributes of users, e.g. a more fine-grained user persona from chit-chat, would be a potential future direction to improve the diversity in simulated user behaviours.

Although LLMs can have better performance on each ToD modelling task and therefore could potentially serve as more powerful modules in EmoLoop, we did not move in this direction since their high computing resource requirement and slow inference speed would hinder their integration into our systems for interactive training and evaluation. Training modular system policy with langEmoUS for 15k dialogues on one Nvidia GeForce RTX 2080 Ti takes around 40 hours. The training time and memory required will be significantly increased if modular systems use LLM-based modules. On the other hand, while LLM-based end-to-end systems may provide a bypass since one LLM is sufficient, implementing RL training on such systems to further leverage task success and emotion signals from the user simulator is another computationally expensive challenge that are yet to solve.

Some of our generative system modules are based on pre-trained language models. Although we have not been reported any harmful generations in the human trail, there is still the possibility for unexpected behaviour when this system is deployed and tested on a very large scale.

For human evaluation, we conducted experiments on Amazon Mechanical Turk platform rather than deployed our systems in the production environment. The participants, despite coming from different countries, are from covering all demographics.

## 9 Ethics Statement

Models, codes and datasets were used in accordance with their respective licenses, terms of use and intended use. The data that we used and generated does not contain any information that names or uniquely identifies individual people or offensive content. The model we used for generating

augmented samples has implemented training objectives for enhanced safety (Appendix C). Systems we used for interaction with real users were very unlikely to generate offensive content as they were fine-tuned on large-scale training data to convey a limited scope of semantic concepts. No offensive content was reported by human users nor observed in post-hoc inspection.

For system conduct annotation, annotators were required to read and agree with our statement of consent for data use before the task. Annotators were paid fairly according to the local regulations of our research institute. We ensured swift communication with annotators so that their concerns were addressed as soon as possible. For poor-quality annotations, we still pay the annotators for their time but block them from our task to ensure data quality and collection efficiency. All annotations are anonymised.

The data annotation and interactive human trial, which involves decision making based on human emotions, have been approved by the ethics review board of the research institute. The proposed system learns how to manipulate human emotional state. Although the system is trained to elicit positive user emotion, this could still be of potential ethical concern and would require greater deliberation when deployed in real-life and more complex scenario.

## 10 Acknowledgement

S. Feng, N. Lubis, and M. Heck are supported by funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research. H-C. Lin and C. van Niekerk are supported by the Ministry of Culture and Science of North Rhine-Westphalia within the framework of the Lamarr Fellow Network. C. Geishauser, B. Ruppik, and R. Vukovic are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018804636). Computing resources were provided by Google Cloud.

## References

Magda B. Arnold. 1960. *Emotion and personality. Vol. I. Psychological aspects*. Columbia Univer. Press.

Trung Bui, Job Zwiens, Mannes Poel, and Anton Ni-

jholt. 2010. *Affective dialogue management using factored pomdps*. *Studies in Computational Intelligence*, 281:207–236.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. *KETOD: Knowledge-enriched task-oriented dialogue*. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Mark Davis. 2018. *Empathy: A Social Psychological Approach*. Routledge, New York.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. *User modeling for spoken dialogue system evaluation*. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 80–87, Santa Barbara, CA, USA. IEEE.

Paul Ekman. 1992. *An argument for basic emotions*. *Cognition & Emotion*, 6:169–200.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. *MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking base-lines*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. *EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gasic. 2023a. *From chatter to matter: Addressing critical steps of emotion recognition learning in task-oriented dialogue*. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 85–103, Prague, Czechia. Association for Computational Linguistics.

- Shutong Feng, Guangzhi Sun, Nurul Lubis, Chao Zhang, and Milica Gašić. 2023b. [Affect recognition in conversations using large language models](#). *CoRR*, abs/2309.12881.
- Christian Geishauer, Carel van Niekerk, Hsien-chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gašić. 2022. [Dynamic dialogue policy for continual reinforcement learning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 266–284, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Christian Geishauer, Carel van Niekerk, Nurul Lubis, Hsien-chin Lin, Michael Heck, Shutong Feng, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2024. [Learning with an open horizon in ever-changing dialogue circumstances](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2352–2366.
- Alicia A. Grandey, Lori S. Goldberg, and S. Douglas Pugh. 2011. [Why and when do stores with satisfied employees have satisfied customers?: The roles of responsiveness and store busyness](#). *Journal of Service Research*, 14(4):397–409.
- Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauer, Hsien-Chin Lin, and Milica Gašić. 2022. [Robust dialogue state tracking with weak supervision and sparse data](#). *Transactions of the Association for Computational Linguistics*, 10:1175–1192.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. [A simple language model for task-oriented dialogue](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.
- Jessica Huynh, Ting-Rui Chiang, Jeffrey Bigham, and Maxine Eskenazi. 2022. [DialCrowd 2.0: A quality-focused dialog system crowdsourcing toolkit](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1256–1263, Marseille, France. European Language Resources Association.
- Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.
- To Eun Kim and Aldo Lipani. 2022. [A multi-task based neural model to simulate users in goal oriented dialogue systems](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, page 2115–2119, New York, NY, USA. Association for Computing Machinery.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.
- Florian Kreyszig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. [Neural user simulation for corpus-based policy optimisation of spoken dialogue systems](#). In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 60–69, Melbourne, Australia. Association for Computational Linguistics.
- Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. [A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning](#). *Machine Intelligence Research*, 20(3):318–334.
- Richard S. Lazarus. 1966. *Psychological stress and the coping process*. McGraw-Hill, New York.
- Esther Levin and Roberto Pieraccini. 1997. [A stochastic model of computer-human interaction for learning dialogue strategies](#). In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, pages 1883–1886, Rhodes, Greece.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. [UniSA: Unified generative framework for sentiment analysis](#). In *Proceedings of the 31st ACM International Conference on Multimedia, MM '23*, page 6132–6142, New York, NY, USA. Association for Computing Machinery.
- Hsien-Chin Lin, Shutong Feng, Christian Geishauer, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2023. [EmoUS: Simulating user emotions in task-oriented dialogues](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2526–2531, New York, NY, USA. Association for Computing Machinery.
- Hsien-chin Lin, Christian Geishauer, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gasic. 2022. [GenTUS: Simulating user behaviour and language in task-oriented dialogues with generative transformers](#). In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–282, Edinburgh, UK. Association for Computational Linguistics.

- Hsien-chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauer, Michael Heck, Shutong Feng, and Milica Gasic. 2021. [Domain-independent user simulation with transformers for task-oriented dialogue systems](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 445–456, Singapore and Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692v1.
- Nurul Lubis, Christian Geishauer, Hsien-chin Lin, Carel van Niekerk, Michael Heck, Shutong Feng, and Milica Gasic. 2022. [Dialogue evaluation with offline reinforcement learning](#). In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 478–489, Edinburgh, UK. Association for Computational Linguistics.
- John Mendonca, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C Farinha, Helena Moniz, Alon Lavie, and Isabel Trancoso. 2023. [Dialogue quality and emotion annotations for customer support conversations](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 9–21, Singapore. Association for Computational Linguistics.
- Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.
- Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiejun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. [Few-shot natural language generation for task-oriented dialog](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.
- Rosalind W. Picard. 2000. *Affective Computing*. The MIT Press, Cambridge, Mass.
- Fuji Ren, Yu Wang, and Changqin Quan. 2015. [TFSM-based dialogue management model framework for affective dialogue systems](#). *IEEJ Transactions on Electrical and Electronic Engineering*, 10(4):404–410.
- Armand Stricker and Patrick Paroubek. 2024. [A Unified Approach to Emotion Detection and Task-Oriented Dialogue Modeling](#). In *IWSDS*, Sapporo (Japan), Japan.
- Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. [Adding chit-chat to enhance task-oriented dialogues](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Carel van Niekerk, Andrey Malinin, Christian Geishauer, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. [Uncertainty measures in neural belief tracking and the effects on dialogue policy performance](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tom Young, Frank Xing, Vlad Pandealea, Jinjie Ni, and E. Cambria. 2021. [Fusing task-oriented and open-domain dialogues in conversational agents](#). In *AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.
- Rui Zhang, Zhenyu Wang, Mengdan Zheng, Yangyang Zhao, and Zhenhua Huang. 2021. [Emotion-sensitive deep dyna-q learning for task-completion dialogue policy learning](#). *Neurocomputing*, 459:122–130.
- Hui Zhu, Xv Wang, Zhenyu Wang, and Kai Xv. 2024. [ESDP: An emotion-sensitive dialogue policy for task-oriented dialogue system](#).
- Qi Zhu, Christian Geishauer, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gasic, and Minlie Huang. 2023. [ConvLab-3: A flexible dialogue system toolkit based on a unified data format](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 106–123, Singapore. Association for Computational Linguistics.

## A EmoWOZ 2.0 Construction

### A.1 Annotation Scheme Justification

Under the framework of the OCC emotion model and the definition of emotional empathy that the observer shares the emotional state of another person (Davis, 2018), we can derive the corresponding emotional response from the system. Considering the following user emotion and situation where:

The user is labelled as *Fearful*, or feeling negative because of an event which has negative consequences on the user his or herself (as defined in EmoWOZ).

An empathetic operator would share the same feeling as the user (therefore also feeling negative). Yet, the feeling in the operator is elicited by an event which has negative consequences on the user (the other party). This feeling is defined as pity, or compassionate in the OCC model.

### A.2 Annotation Interface

We adopted the same annotation set-up, annotator selection criteria, and quality assurance approaches as outlined by Feng et al. (2022). Each utterance is annotated by three annotators, who were provided with the entire preceding dialogue history when annotating the current utterance. Annotators were English speakers. The final label was obtained from majority voting. When the agreement could not be reached, a fourth annotator was introduced. Overall, 54 crowd workers have contributed to our study.

Operator Conduct Annotation	
<b>Instructions (Click to collapse)</b>	
In this task, you will be shown dialogue turns between a <b>User</b> and an <b>Operator</b> . You will be supplied with the following information:	
<ul style="list-style-type: none"> <li>User's request</li> <li>Operator's response</li> </ul>	
You will then be prompted to answer the following questions regarding the <b>Operator</b> :	
Which of the following best describes the conduct of the <b>Operator</b> ?	
<b>Compassionate</b>	The operator is showing sympathy, concern, or sadness for the user (e.g. because of the user's suffering or bad luck), and is wanting to help the user. E.g. <b>User:</b> I appreciate your help. Do you know how quickly the police will respond? I may need some medical help as well. <b>Operator:</b> The police will arrive soon, are you injured?
<b>Apologetic</b>	The operator is expressing regretful acknowledgement of an offence or failure. E.g. <b>User:</b> Yes, please. 6 people 3 nights starting on Tuesday. <b>Operator:</b> I am sorry but I wasn't able to book that for you for Tuesday. Is there another day you would like to stay or perhaps a shorter stay?
<b>Enthusiastic</b>	The operator is showing extra helpfulness (e.g. using emotionally colourful words when trying to convince the user about a recommendation), or feeling happy for the user because of something good happening to the user. <b>User:</b> Are there anything fun to do in city centre? <b>Operator:</b> My favorite attraction in the centre of the city is a concert hall called Man on the Moon. It is amazing! They are at 2 Norfolk Street.
<b>Appreciative</b>	The operator is showing pleasure for successfully fulfilling the user's goal or request. <b>User:</b> You were great. Goodbye. <b>Operator:</b> We are happy to help. Have a good day!
<b>Neutral</b>	The operator does not explicitly show any emotional conduct. E.g. <b>User:</b> I am departing from Birmingham New Street. <b>Operator:</b> Can you confirm your desired travel day?
If you have any questions, please contact: [REDACTED]	
Question #1 (PMUL4981.json)	
<b>User:</b> Hey there, I want to get a train to Cambridge for Thursday please	
<b>Operator:</b> Okay, from where are you departing?	
Which of the following best describes the conduct of the <b>Operator</b> ?	
<input type="radio"/> <b>Compassionate</b> (The operator is showing sympathy, concern, or sadness for the user (e.g. because of the user's suffering or bad luck), and is wanting to help the user.) <input type="radio"/> <b>Apologetic</b> (The operator is expressing regretful acknowledgement of an offence or failure.) <input type="radio"/> <b>Enthusiastic</b> (The operator is showing extra helpfulness (e.g. using emotionally colourful words when trying to convince the user about a recommendation), or feeling happy for the user because of something good happening to the user.) <input type="radio"/> <b>Appreciative</b> (The operator is showing pleasure for successfully fulfilling the user's goal or request.) <input type="radio"/> <b>Neutral</b> (The operator does not explicitly show any emotional conduct.)	
Question #2 (PMUL4981.json)	
<b>User:</b> I am coming in from the Stansted Airport.	
<b>Operator:</b> What time will you need to be traveling?	
Which of the following best describes the conduct of the <b>Operator</b> ?	
<input type="radio"/> <b>Compassionate</b> (The operator is showing sympathy, concern, or sadness for the user (e.g. because of the user's suffering or bad luck), and is wanting to help the user.) <input type="radio"/> <b>Apologetic</b> (The operator is expressing regretful acknowledgement of an offence or failure.) <input type="radio"/> <b>Enthusiastic</b> (The operator is showing extra helpfulness (e.g. using emotionally colourful words when trying to convince the user about a recommendation), or feeling happy for the user because of something good happening to the user.) <input type="radio"/> <b>Appreciative</b> (The operator is showing pleasure for successfully fulfilling the user's goal or request.) <input type="radio"/> <b>Neutral</b> (The operator does not explicitly show any emotional conduct.)	

Figure A.1: Web-interface for conduct annotation.

### A.3 Annotator Confusion Matrix

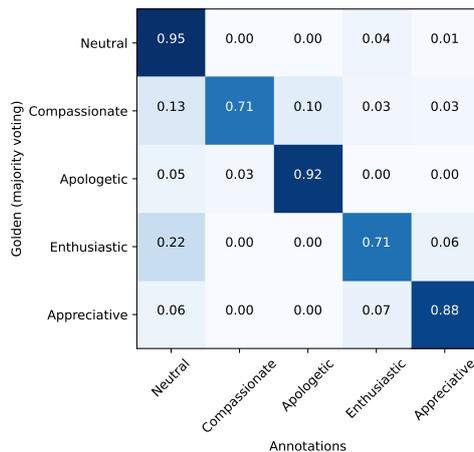


Figure A.2: Annotator confusion matrix.

### A.4 System Conduct Distribution

Conduct	Count	Proportion
Neutral	52,236	73.0%
Appreciative	9,763	13.6%
Enthusiastic	6,364	8.9%
Apologetic	3,049	4.3%
Compassionate	112	0.2%

Table A1: Conduct distribution in MultiWOZ.

## B User Simulator Implementation Details

Following the setting in Lin et al. (2023), the input and output of langEmoUS are represented as JSON-formatted strings, which are composed of tokens in natural language. We initialised our model based on the BART model (Lewis et al., 2020) and fine-tuned it on our EmoWOZ 2.0 dataset. We optimised our model with Adam (Kingma and Ba, 2015), where the learning rate is  $2e^{-5}$  for 5 epochs. As shown in Table B1, langEmoUS achieves state-of-the-art performance on user sentiment and emotion prediction.

Model	Sentiment	Emotion
SatActUtt (Kim and Lipani, 2022)	0.379	-
EmoUS (Lin et al., 2023)	0.693	0.501
langEmoUS	<b>0.742</b>	<b>0.521</b>

Table B1: Performance for emotion and sentiment prediction of different models by measuring macro-F1 score.

## C Natural Language Generator Implementation Details

### C.1 NLG Training

#### C.1.1 Training Configuration

We trained SC-BART and SEC-BART on EmoWOZ 2.0. We trained our model with Adam optimiser for standard cross entropy loss where the learning rate was set to  $2e^{-5}$  for 5 epochs (with an early-stopping criterion based on the loss in the validation set) and a batch size of 16. During inference, we set the temperature to 0.9 and a beam number of 2 to promote some degree of diversity.

### C.1.2 Prompt Template

Our NLG models take the following input: previous user utterance  $u_t$ , dialogue semantic actions  $a_t$ , and conduct  $e_t^{sys}$  (for SEC-BART only). The prompt template is shown as follows:

**SEC-BART** Given the previous user request “ $\{u_t\}$ ”, the natural language realisation of dialogue action “ $\{a_t\}$ ” with a/an “ $\{e_t^{sys}\}$ ” conduct is

**SC-BART** Given the previous user request “ $\{u_t\}$ ”, the natural language realisation of dialogue action “ $\{a_t\}$ ” is

Given the prompt, the model predicted the probability distribution for a sequence of tokens. The output target is the corresponding ground-truth system response in EmoWOZ 2.0.

### C.1.3 Model Performance

Model	BLEU $\uparrow$	SER $\downarrow$
SC-GPT (Peng et al., 2020)	33.6	4.8
T5NLG (Zhu et al., 2023)	35.8	3.7
SC-BART	<b>35.9</b>	3.9
SEC-BART	34.9	<b>3.6</b>

Table C1: NLG Performance.

## C.2 Data Augmentation

### C.2.1 Augmented Sample Collection

Since the conduct distribution in EmoWOZ 2.0 is heavily imbalanced, we leveraged large language models for data augmentation. We selected system utterances with neutral conduct as the source to paraphrase for a target non-neutral conduct. We used LLaMA-2-13b-chat model (Touvron et al., 2023). We used the following prompt:

Given the user request “ $\{u_t^{usr}\}$ ” and the operator response action “ $\{a_t\}$ ”, please paraphrase the operator response “ $\{u_{t,groundtruth}^{sys}\}$ ” in a more “ $\{e_{t,target}^{sys}\}$ ” way? Please only give the answer, in less than  $2 \times len(u_{t,groundtruth}^{sys})$  tokens and enclosed with [RESP][[/RESP].

We also experimented with ICL but the model tends to over-fit on the ICL samples. We therefore let it paraphrase in an zero-shot set-up to best explore its knowledge from pre-training for better diversity in the expression.

### C.2.2 Augmented Sample Selection

Since the model does not always follow the target conduct. For example, the large language model (LLM) would find some action-conduct combinations unreasonable. We therefore applied filtering on the LLM-generated samples.

**Conduct Expressiveness** We trained an ensemble of 10 ContextBERT-ERToD models for conduct classification on EmoWOZ 2.0. The classifier reports an average weighted F1 score of 81.8% without neutral. We then used majority voting from the classifier ensemble to correct the original target conduct when generating the sample.

**Faithfulness to Semantic Action** We used the rule-based script in ConvLab-3 to evaluate NLG slot error rates in the paraphrased output based on the dialogue actions in the prompt. If there are slot errors in the output, we drop the sample.

Overall, we obtained 949 samples for *Compassionate*, 900 for *Apologetic*, 2274 for *Enthusiastic*, and 490 for *Appreciative*.

## D Human Evaluation

### D.1 Web Interface

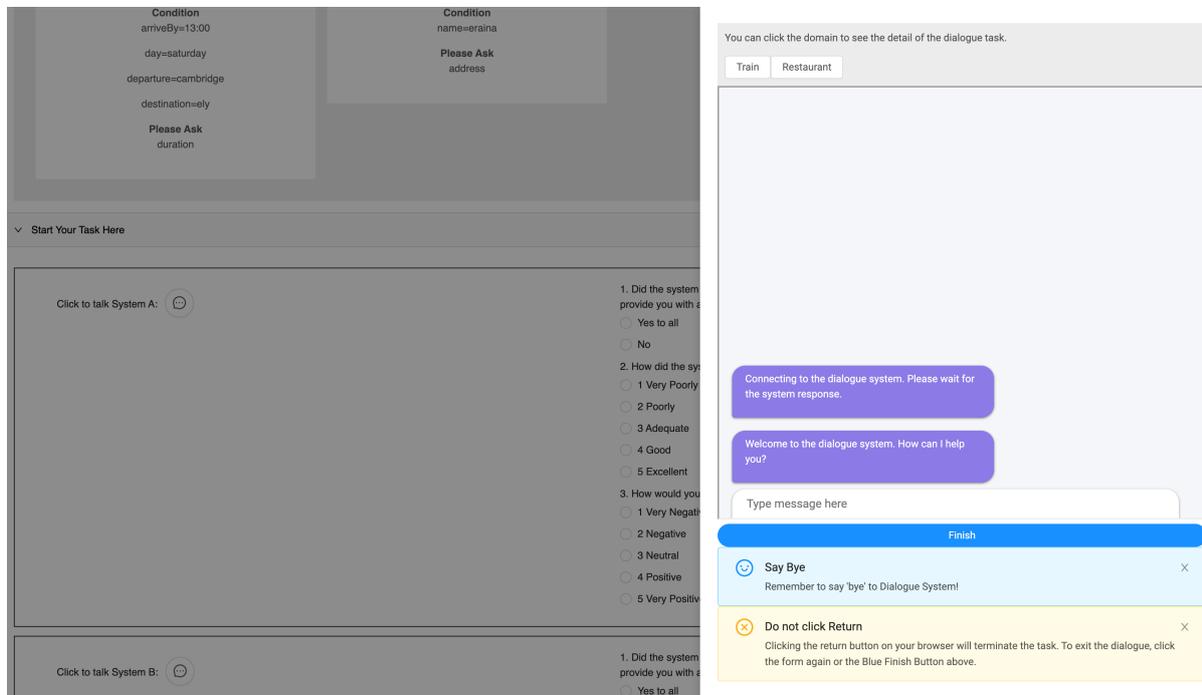


Figure D.1: The web interface for human trial.

### D.2 Survey Questions

#### D.2.1 Question 1 - Task Success

**Question** Did the system find what you look for? Did it provide all the information that you need? If you ask for a booking, did it provide you with a reference number?

**Multiple Choices** (A) Yes to all; (B) No.

#### D.2.2 Question 2 - Sentiment Rating

**Question** How would you rate your sentiment after the conversation?

**Multiple Choices** (A) Very Negative; (B) Negative; (C) Neutral; (D) Positive; (E) Very Positive.

### D.3 Dialogue Excerpts

System	Role	Text
EmoLoop	User	Hello, I am looking for an attraction called byard art. <i>I am really excited about my trip to Cambridge!!</i>
	System	I'm sorry, there is byard art in the south. The address is 14 kings parade.
	User	<i>Why are you apologising?</i>
	System	<i>I apologize for the confusion.</i> If you need to contact them, please call 01223464646
SimpleLoop	User	No, I just want to book a hotel for 3 people and 3 nights on friday and I need the address.
	System	I can do that. What is the name of the hotel?
	User	Can you suggest one?
	System	I have booked a taxi for you. It is a toyota and the contact number is 29929243243
	User	<i>Sorry, you are not helping at all</i>
	System	I was able to book that for you. The reference number is 63484139414. What is your destination?

Table D1: Human trial dialogue excerpts from EmoLoop and SimpleLoop in a similar situation where both systems were not performing adequately at the beginning. EmoLoop responded with an improper conduct but realised the mistake and continued to provide information. SimpleLoop did not show any sign of realising the mistake.

System	Role	Text
EmoLoop	User	I need the phone number of the hotel with name ashley hotel.
	System	<i>Sure thing!</i> It is located in the north and the postcode is cb41er and the phone number is 01223350059.
	User	<i>Thanks a lot.</i> I also need a spanish restaurant with moderate price.
	System	I would recommend la tasca. It is located at 14-16 bridge street.
SimpleLoop	User	<i>Alright, I am also interested</i> in a museum in the city centre.
	System	I recommend broughton house gallery. It is a museum. Their phone number is 01223314960.
	User	<i>Nice,</i> could you also give me the address and postcode?
System	The postcode is cb11ln and the phone number is 01223314960.	

Table D2: Human trial dialogue excerpts from EmoLoop and SimpleLoop in a similar situation where the user expressed excitement and appreciation. Both systems performed adequately in terms of completing the user's goal. EmoLoop responded in an enthusiastic way whereas SimpleLoop did not respond with affective conduct.

System	Role	Text
EmoLoop	User	I need some information.
	System	What kind of information do you need?
	User	Which department there?
	System	<i>I am sorry, I don't understand. What do you mean?</i>
	User	I need information about haematology.
	System	The Addenbrookes Hospital is located at Hills Rd, Cambridge , postcode cb20qq. The phone number is 01223216297.
SimpleLoop	User	I need information about haematology day unit.
	System	The pizza hut cherry hilton Hospital is located at Hills Rd, Cambridge , cb20qq. The phone number is 01223216297.
	User	Department details?
	System	There are 10 trains that go to london kings cross on sunday. The TR7423 leaves at 05:00 and arrives at 05:51. Would you like me to book that for 6 people?

Table D3: Human trial dialogue excerpts from EmoLLAMA and SimpleLLAMA in a similar situation where the user asked for hospital information. EmoLLAMA, although not understanding the user's vague request, attempted to apologise and clarify the request with the user. The task was successfully completed. SimpleLLAMA carried on with errors and fails the task. (The wrong values, "pizza hut" and "cherry hilton", in the SimpleLLAMA response were due to wrong domains in the dialogue action prediction, which led to corresponding name slots in the response. These name slots were filled during lexicalisation based on actions, resulting in an obviously irrational output.)

System	Role	Text
EmoLoop	User	Recommend a taxi which arrives by 4:15 pm from the fez club to tandoori place.
	System	I'd be happy to help with your request, first I'll need to know where you'll be leaving from, where you're going, and what time you'd like to arrive.
	User	I will be leaving from fez club and going to tandoori palace arriving by 4:15 pm.
	System	Booking was successful. The car type is bmw and the contact number is 19287599612.
SimpleLoop	User	I want to book a taxi for 16:15 from cherry hinton water play to ely train station.
	System	What is your destination?
	User	Ely train station.
	System	Booking completed. Booked car type: ford, Contact number: 46563428252.

Table D4: Human trial dialogue excerpts from EmoLLAMA and SimpleLLAMA in a similar situation where both systems failed to capture all information provided in the user request. EmoLLAMA at first missed the information provided by the user but replied in a compassionate way. The user repeated and then the system provides the correct information. Likewise, SimpleLLAMA missed the destination in the first turn. After the user repeated, the system completed the task for the user. Yet, there is no affective interaction between the user and SimpleLLAMA.

## E Further Analysis

### E.1 Ablation Study for EmoLoop with Supervised Training Only

System	Und	Gen	Man	Success	Sentiment
SimpleLoop-SL	-	-	-	0.512	0.244
	+	-	-	0.494	0.246
	-	+	-	0.493	0.249
EmoLoop-SL	+	+	-	0.516	0.273

Table E1: Success and average user sentiment of our system variants from the interactive evaluation with langEmoUS. +/- means whether the emotion is involved in the corresponding ToD stage: **U**nderstanding, **M**anagement, or **G**eneration. All systems are trained via SL.

### E.2 Impact of Training Set-ups on System Conduct

We investigate how the EmoLoop's affective behaviour is shaped in different stages of training. Figure E.1 shows the distribution of system conduct at different dialogue turns in EmoWOZ 2.0, and policy output during interaction with langEmoUS after supervised pre-training and language-level RL. Comparing Figure E.1a and Figure E.1b suggests that the policy imitates the affective behaviour of operators in the corpus.

After RL, the policy is more inclined to express *enthusiastic* and *appreciative* while expressing *compassionate* and *apologetic* less frequently. This illustrates the affective strategy of the policy to elicit more positive emotions in the simulated user.

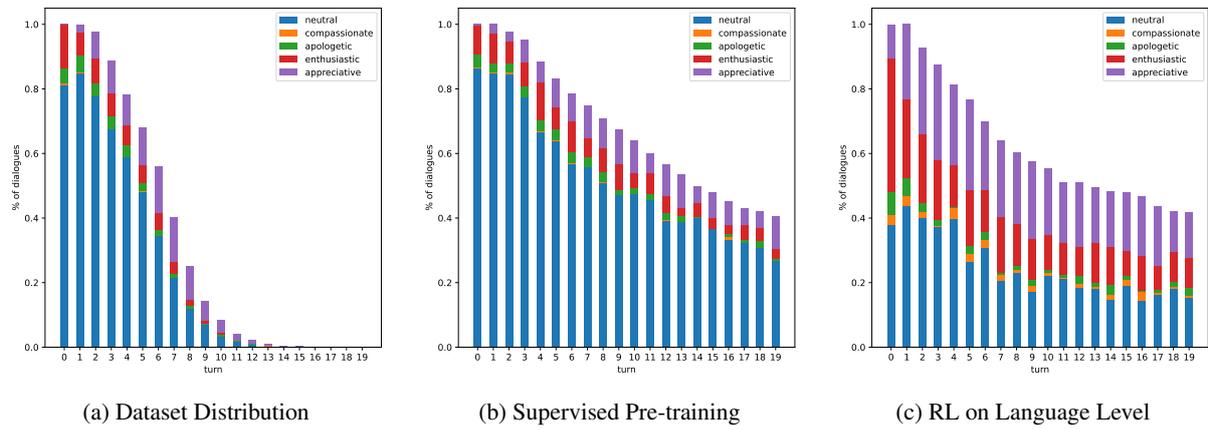


Figure E.1: Distributions of system conduct for different turn positions at different stages of policy training.



## Chapter 9

# Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation

This chapter summarises our work on designing and optimising emotionally intelligent task-oriented dialogue systems using reinforcement learning with large language models, and gives a verbatim copy of our manuscript (Feng et al., 2025):

Shutong Feng et al. (2025). *Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation*. arXiv: 2507.01594 [cs.CL]. URL: <https://arxiv.org/abs/2507.01594>

### 9.1 Summary

In Chapter 8, we build emotion-aware task-oriented dialogue systems under modular and end-to-end architectures. While this prior work affirms the feasibility of emotion-aware ToD systems, the optimal choice of system architectures, representations of information, and optimisation strategies is still unexplored.

In the work presented in this chapter, we systematically explore a wide range of design considerations. We examine systems under various architectural, representational, and optimisation design considerations for developing emotionally intelligent ToD systems in noisy, ambiguous conversational settings. In particular, we explore a previously unexplored combination of design choices and build an end-to-end LLM-based and emotionally intelligent task-oriented dialogue system trained via reinforcement learning. We name the model **LUSTER**, LLM-based **Unified System for Task-oriented dialogue with End-to-end Reinforcement learning**. LUSTER employs fully lexicalised representations and leverages both short-term rewards reflecting user sentiment and long-term rewards reflecting task success, enabling the system to adapt dynamically to users' emotional signals while still fulfilling task objectives.

We evaluate systems with a natural-language user simulator and a noisy natural language understanding module, simulating real-world interaction challenges. Experiments show that combining LLM capabilities with structured reward modelling leads to more resilient, emotionally responsive dialogue agents, offering a promising direction for the next generation of task-oriented systems.

### 9.2 Personal Contributions

The implementation, technical results, and writing are my own work. Hsien-Chin Lin provided the user simulation results of the agentic system. Nurul Lubis contributed to the design of the reinforcement learning algorithm. Milica Gasić supervised the project. Co-authors assisted in writing and proofreading.

# Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation

Shutong Feng, Hsien-chin Lin, Nurul Lubis, Carel van Niekerk, Michael Heck,  
Benjamin Ruppik, Renato Vukovic, Milica Gašić

Heinrich Heine University Düsseldorf, Germany

{fengs, linh, lubis, niekerk, heckmi, ruppik, revuk100, gasic}@hhu.de

## Abstract

Task-oriented dialogue (ToD) systems are designed to help users achieve specific goals through natural language interaction. While recent advances in large language models (LLMs) have significantly improved linguistic fluency and contextual understanding, building effective and emotionally intelligent ToD systems remains a complex challenge. Effective ToD systems must optimise for task success, emotional understanding and responsiveness, and precise information conveyance, all within inherently noisy and ambiguous conversational environments. In this work, we investigate architectural, representational, optimisation as well as emotional considerations of ToD systems. We set up systems covering these design considerations with a challenging evaluation environment composed of a natural-language user simulator coupled with an imperfect natural language understanding module. We propose **LUSTER**, an LLM-based Unified System for Task-oriented dialogue with End-to-end Reinforcement learning with both short-term (user sentiment) and long-term (task success) rewards. Our findings demonstrate that combining LLM capability with structured reward modelling leads to more resilient and emotionally responsive ToD systems, offering a practical path forward for next-generation conversational agents.

## 1 Introduction

Conversational artificial intelligence (AI) enables machines to engage in natural, human-like interactions using spoken or written language. The rapid advancement of large language models (LLMs) has significantly enhanced the capabilities of conversational AI (Touvron et al., 2023; Abdin et al., 2024). Trained on vast amounts of text data, LLMs can generate fluent, context-aware responses, and

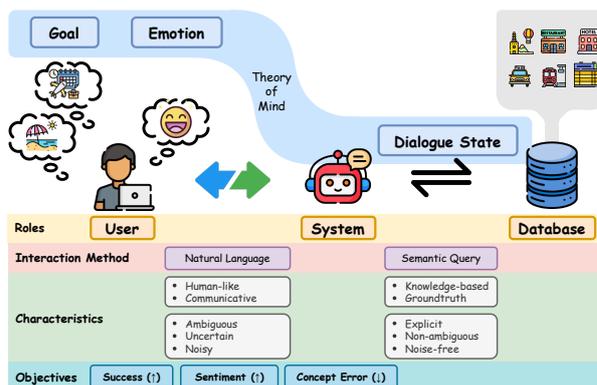


Figure 1: Goal-driven Human-machine Interaction

follow complex instructions. This significantly improves chat applications in both linguistic fluency and handling of nuanced or open-ended interactions (OpenAI, 2023; DeepSeek-AI, 2025).

Among the diverse conversational AI applications, task-oriented dialogue (ToD) systems stand out as a specialised class to help users accomplish a goal through several turns of natural language interactions (Jurafsky and Martin, 2025) (see Figure 1). Beyond task completion, effective communication also hinges on the user’s emotional experience (Picard, 1997). The user expresses both goal-driven and emotional cues in inherently ambiguous, uncertain, and noisy natural language (Feng et al., 2022). The system interprets this input and estimates a dialogue state, an internal representation of the user’s goal, functionally akin to a theory of mind (Premack and Woodruff, 1978). This state, typically represented symbolically (e.g., slot-value pairs), supports precise, unambiguous queries to a database, which likewise stores information in a structured, symbolic format. Thus, high-performing ToD systems must (1) **maximise task success** by correctly fulfilling the user’s goal, (2) **enhance sentiment** by maintaining a positive emotional interaction, and (3) **minimise error** in conveying database information. Related errors are termed hallucina-

tions in natural language generation (NLG) (Jurafsky and Martin, 2025), word errors in automatic speech recognition (Young et al., 2006), and reasoning errors in neural-symbolic systems (Garcez et al., 2002). In ToDs, explicitly represented semantics enable detection of both missing and hallucinated values, and we collectively term them as *concept errors*.

Designing a ToD system involves critical design and training choices that shape its ability to understand and assist users. A key **architectural** choice lies in whether to adopt a modular, end-to-end, or agentic approach. Modular systems separate the dialogue pipeline into components such as natural language understanding (NLU), dialogue state tracking (DST), policy, and NLG, each trained individually (Young et al., 2013). In contrast, end-to-end systems model the entire pipeline as a single unit for reduced error propagation and joint optimisation (Wen et al., 2017). Recently, agentic designs have emerged, leveraging LLMs’ reasoning capability for task-oriented decision-making (Yao et al., 2023).

These architectures align with different **optimisation strategies**. The traditional divide-and-conquer method trains each module in isolation. While this approach offers a certain level of explainability of intermediate processing steps, there could be information loss between components, leading to overall suboptimal behaviour. Alternatively, optimising the dialogue policy in the full interactive loop can better adapt decision-making to natural-language interaction outcomes (Feng et al., 2024a). End-to-end trainable systems, if paired with reinforcement learning (RL), can allow holistic user adaptation but face challenges like data inefficiency and the need for sophisticated reward design.

**Emotion modeling** is another essential yet often overlooked aspect of ToDs (Feng et al., 2022). Systems may handle emotion implicitly via hidden language representations, or explicitly through dedicated emotion detection and response planning stages (Stricker and Paroubek, 2024). In RL, emotion can also be integrated into training by using user sentiment as a short-term reward to guide policy updates (Feng et al., 2024a).

The representational considerations vary across systems. For the **dialogue state**, most systems use explicit tracking to provide clear intermediate representations for downstream tasks and database

query, though latent states are also possible (Lin et al., 2024). **Dialogue actions** can likewise be explicit (Geishhauser et al., 2022; Hosseini-Asl et al., 2020), typically delexicalised (domain, intent, slot types but no values), or implicit, encoded as latent vectors learned by the model (Lubis et al., 2020). While modular systems use NLG to produce lexicalised responses, end-to-end systems usually output delexicalised **responses** to facilitate learning and evaluation. However, the strong lexical capabilities of modern LLMs are reducing the necessity of delexicalisation in both action planning and response generation.

Finally, **training methodology** critically impacts system performance. Supervised learning (SL) on annotated data offers a strong foundation but struggles with diverse user behaviours. RL enables adaptive optimisation using short-term (e.g., user sentiment) and long-term (e.g., task success) rewards from simulated or real-world interactions. A hybrid training paradigm that combines SL with RL often best balances stability and adaptability (Williams et al., 2017).

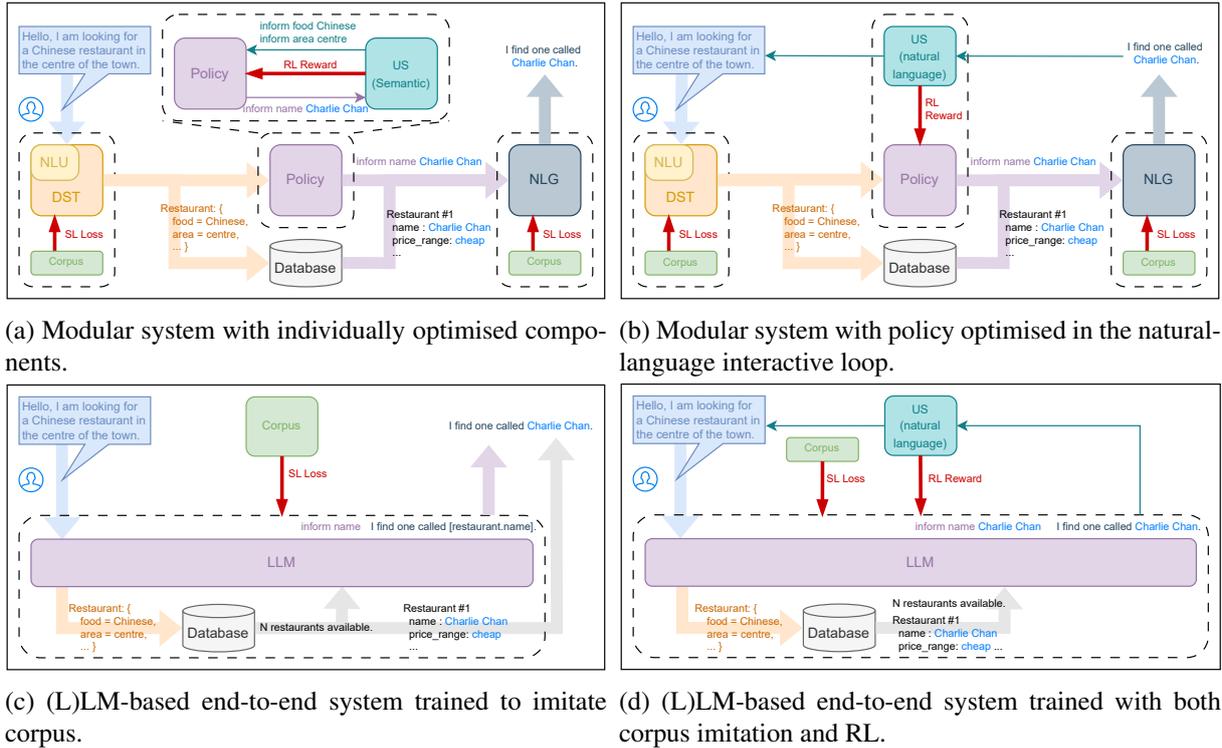
Together, these design axes form a rich space of possibilities and trade-offs involving scalability, flexibility, interpretability, and robustness. Our contributions are threefold:

- We construct a diverse suite of ToD systems representative of each design and training paradigm (Figure 2), and systematically analyse them.
- We harness the advantages of different design considerations in a principled manner and address a previously unexplored combination of design choices. Specifically, we build the first end-to-end ToD system that integrates fully lexicalised representations, an LLM backbone, and RL with both short-term affective and long-term task success rewards via online user simulation.
- With our proposed method, we build **LUSTER**, an LLM-based Unified System for Task-oriented dialogues with End-to-end Reinforcement learning. LUSTER substantially improves task success and reduces concept errors compared to alternative approaches.

## 2 Related Work

### 2.1 Modular Systems

Traditional ToD systems typically follow a modular pipeline, with separate components: NLU for user intent understanding, DST for user goal estimation, policy for system action selection, and



(a) Modular system with individually optimised components.

(b) Modular system with policy optimised in the natural-language interactive loop.

(c) (L)LM-based end-to-end system trained to imitate corpus.

(d) (L)LM-based end-to-end system trained with both corpus imitation and RL.

Figure 2: ToD system training paradigms. Red arrows indicate learning signals for components in dotted boxes.

NLG for converting actions into natural utterances. Systems also interface with a database (DB) and rely on an ontology to structure symbolic queries and information abstraction.

This divide-and-conquer architecture (Figure 2a) ensures clear separation of objectives and allows component-level supervision. Modules are typically trained independently on annotated datasets such as MultiWOZ (Budzianowski et al., 2018), with SL applied to NLU, DST, and NLG, and RL applied to the policy. The interaction is framed as a partially observable Markov decision process (POMDP) (Young et al., 2013), and policy learning is carried out with simulated (Schatzmann et al., 2007; Kreyssig et al., 2018; Lin et al., 2023) or real (Gašić et al., 2013) users.

However, policies trained in isolation often misalign with other pipeline components during full-system interaction. To mitigate this issue, Feng et al. (2024a) optimise the policy within the natural-language interactive loop (Figure 2b).

To manage the complexity of the policy output space, modular dialogue policies typically produce delexicalised dialogue actions, omitting actual slot values. This abstraction simplifies policy learning by reducing action variability. The NLG module then realises these actions, together with slot values from database query results, into natu-

ral language utterances (Figures 2a and 2b).

## 2.2 End-to-end Systems

To overcome challenges in information passing and module coordination of modular pipelines, researchers have developed end-to-end systems that unify learning and eliminate the need for information transmission between modules. Early approaches demonstrate the feasibility of training dialogue systems as a single monolithic model (Wen et al., 2017). With advances in language models (LMs), researchers (Hosseini-Asl et al., 2020; Peng et al., 2021; Stricker and Paroubek, 2024) use decoder-only LMs to map dialogue history to belief states, database queries, and final responses in an auto-regressive manner (Figure 2c).

End-to-end systems typically retain the practice of using delexicalised dialogue actions from modular pipelines. They also generate responses containing placeholders (e.g.,  $\langle v . food \rangle$ ) rather than actual entity values. This allows weight sharing across values of the same slot type and facilitates corpus-based supervised learning and evaluation. The substitution of placeholders with retrieved slot values is handled as a separate, post-processing operation (Figure 2c).

While these models perform competitively on datasets, the lexicalisation of these responses is

not robust and often leads to failures in interactive settings (Feng et al., 2024a).

Several efforts have incorporated RL into the training of end-to-end ToD systems. Jang et al. (2022) combine offline behavior cloning with a critic model that evaluates and filters self-generated delexicalised responses from an LM-based end-to-end ToD system. Lubis et al. (2020) tackle the challenge of the large action space inherent in word-level sequential decision-making by introducing a variational model that defines a latent action space. This abstraction enables more tractable RL training via simulated yet still delexicalised interactions for an end-to-end ToD system.

### 2.3 Agentic Systems

The advent of general-purpose LLMs has enabled agentic ToD systems that perform complex dialogue reasoning via prompting rather than task-specific training. These models adapt to diverse dialogue tasks via prompt engineering using task descriptions, demonstrations, or dialogue history under a zero- or few-shot setting. Techniques such as prompt chaining (Wei et al., 2022) and self-refinement loops (Yao et al., 2023) enhance reasoning and consistency, though at the cost of higher computational complexity.

ToD system developers have leveraged this ability to construct both individual components, such as slot filler (Sun et al., 2024), dialogue state tracker (Heck et al., 2023), and emotion recognition (Feng et al., 2024b), as well as agentic systems using carefully designed prompts (Li et al., 2024). Vukovic et al. (2024) demonstrates that chain-of-thought prompting can also be used to automate ontology construction, significantly reducing human effort in domain specification.

Despite their flexibility, agentic systems come with trade-offs. LLMs operating in zero- or few-shot settings remain sensitive to prompt phrasing, often exhibit unstable behaviour across sessions, and lack consistent control over generated outputs (Errica et al., 2025). They also underperform compared to models with task-specific optimisation, particularly in high-stakes or domain-intensive scenarios where robustness and consistency are critical (Hudeček and Dusek, 2023).

**In summary**, modular systems offer inter-pretability but suffer from information loss across modules; end-to-end systems provide unified learning but face challenges in out-of-corpus gen-

eralisation; agentic approaches offer flexibility and low training overhead but still lag in reliability and controllability. Their complementary strengths motivate a unified approach explored in this work to combine all potential benefits.

## 3 LUSTER: LLM-based Unified System for Task-oriented dialogue with End-to-end Reinforcement Learning

### 3.1 Inference with LUSTER

As illustrated in Figure 3, LUSTER performs sequential inference over core ToD subtasks, appending each output to the prompt to guide the next generation step. This auto-regressive process enables dynamic integration of contextual and affective cues. Motivated by prior work in affective modelling in ToDs (Feng et al., 2024a), LUSTER augments traditional ToD subtasks (DST, policy optimisation, and NLG) with two additional tasks: user emotion recognition and system conduct (affective behaviour) selection. This supports emotionally intelligent and goal-aligned responses.

Unlike prior (L)LM-based end-to-end systems that separate generation steps with special tokens (e.g., `<|state|>`), LUSTER adopts a lightweight prefix-based format inspired by Li and Liang (2021). Each task starts with a natural-language cue (e.g., “state :”) and ends with a standard end-of-sequence token. This simplifies implementation and improves compatibility with LLM tokenisers. We also explored rare or reserved tokens as markers, but they are often skipped or misgenerated due to LoRA-induced numerical imprecision, destabilising downstream behaviour.

**Step 0: Dialogue History Encoding** At the  $t$ -th turn, LUSTER receives a dialogue history  $\mathcal{H}_t$ , constructed as an alternating sequence of user and system utterances, concluding with the most recent user input:

$$\mathcal{H}_t := \{\text{user} : u_{\text{usr},1}, \text{system} : u_{\text{sys},1}, \dots, \text{user} : u_{\text{usr},t}\} \quad (1)$$

**Step 1: User Emotion Recognition** The model first predicts the user’s emotional state  $e_t$  by selecting the token from a discrete set of emotions  $E$  corresponding to the highest logit score:

$$e_t = \arg \max_{e \in E} \text{LLM}(\mathcal{H}_t, \text{“emotion :”}).\text{logit}(e) \quad (2)$$

**Step 2: Active Domain Recognition** Next, the model identifies the active domain  $d_t$  from the pre-defined set of domains  $D$ :

$$d_t = \arg \max_{d \in D} \text{LLM}(\mathcal{H}_t, e_t, \text{“domain :”}).\text{logit}(d) \quad (3)$$

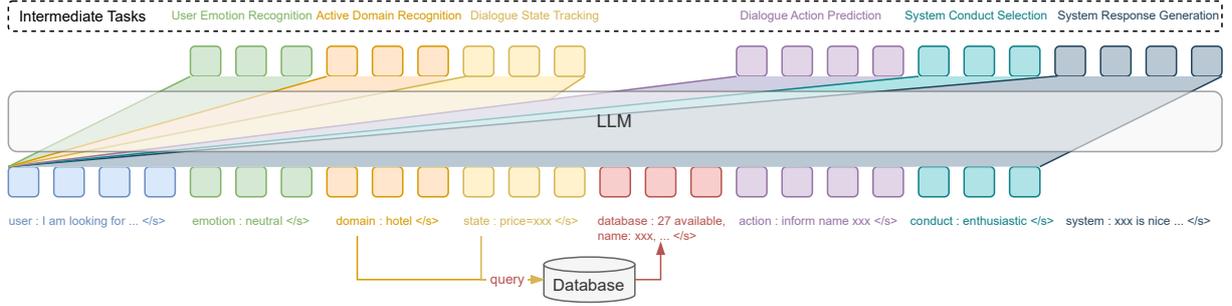


Figure 3: LUSTER model architecture and supervised training objectives.

**Step 3: Dialogue State Tracking** The dialogue state  $ds_t$  is generated for the selected domain only, in order to limit memory usage and maintain relevance. All  $m$  slots of the active domain are filled according to the ontology with either: 1) a freely generated value for non-categorical slots, 2) a selected value from a candidate list for categorical slots, or 3) “unknown” if the slot is not mentioned by the user.

$$ds_t := \{\text{slot}_1 : \text{value}_1; \dots; \text{slot}_m : \text{value}_m\} \quad (4)$$

$$ds_t = \text{LLM}(\mathcal{H}_t, e_t, d_t, \text{“state :”}) \quad (5)$$

**Step 4: Database Query** The generated state is then used to retrieve the list of matching entries  $ent$  from the database. The database description  $db_t$  includes the number of matching entities and all attribute-value pairs of the first matched entity:

$$ent = \text{database.query}(d_t, ds_t) \quad (6)$$

$$db_t := \{\text{len}(ent) \text{ found} : \text{attr val} \forall (\text{attr}, \text{val}) \in ent[0]\} \quad (7)$$

**Step 5: Dialogue Action Prediction** Next, the model generates dialogue actions  $da_t$  as a set of intent-slot-value tuples:

$$da_t := \{\text{intent}_1 \text{ slot}_1 \text{ value}_1; \dots\} \quad (8)$$

$$da_t = \text{LLM}(\mathcal{H}_t, e_t, d_t, ds_t, db_t, \text{“action :”}) \quad (9)$$

**Step 6: System Conduct Selection** The system’s emotional stance  $c_t$  is then selected from a set of conducts  $C$ :

$$c_t = \arg \max_{c \in C} \text{LLM}(\mathcal{H}_t, e_t, d_t, ds_t, db_t, da_t, \text{“conduct :”}).\text{logit}(c) \quad (10)$$

**Step 7: System Response Generation** Finally, LUSTER generates the system’s natural language response  $u_{\text{sys},t}$  conditioned on the full context:

$$u_{\text{sys},t} = \text{LLM}(\mathcal{H}_t, e_t, d_t, ds_t, db_t, da_t, c_t, \text{“system :”}) \quad (11)$$

### 3.2 Fully Lexicalised Representation

Different from how lexicalisation is involved in the generation process of modular and end-to-end

systems as discussed in Section 2.1 and 2.2, LUSTER uses fully lexicalised representation throughout the generation, leveraging the world knowledge and generative strength of large pretrained LLMs (Figure 2d and 3): (1) **Dialogue actions** include slot values along with intents and slots. (2) **Responses** are generated directly in lexicalised form, eliminating placeholders and subsequent substitution. (3) The **database results** are represented as structured, value-rich entity descriptions, not just query summaries (e.g., “3 options found”).

LUSTER’s lexicalised approach simplifies inference by eliminating the need for post-hoc lexicalisation. By using fully lexicalised representations and natural-language prefixes (e.g., state:”), we avoid modifying the LLM’s vocabulary or resizing its embedding and output layers. This means these layers do not need to be fully re-trained, even when using low-rank adaptation (LoRA Hu et al. 2022). As a result, the number of trainable parameters is significantly reduced, making RL via online interaction feasible. Moreover, the lexicalised setup enhances modularity and future compatibility, enabling easy substitution of the LLM backbone as newer models emerge.

### 3.3 Training Objectives

#### 3.3.1 Supervised Learning

An overview of LUSTER’s training procedure is shown in Figure 2d. We begin by training the model to imitate successful demonstrations from human operators recorded in the EmoWOZ corpus (Feng et al., 2022). Each training example is formatted from a dialogue turn as a fully lexicalised sequence  $x = \{\mathcal{H}_t, e_t, d_t, ds_t, db_t, da_t, c_t, u_{\text{sys},t}\}$ , following the inference structure described in Section 3.1. Here,  $x$  is a token sequence of length  $n = |x|$ , denoted by  $(x_1, x_2, \dots, x_n)$ .

We train LUSTER using the standard causal language modelling objective (Radford et al.,

2019). Given a dataset  $\mathcal{X} = \{x^1, x^2, \dots, x^{|\mathcal{X}|}\}$ , we optimise the model parameters  $\theta$  to minimise the negative log-likelihood of the tokens in each sequence:

$$p(x) = \prod_{i=1}^n p(x_i | x_{<i}), \quad \mathcal{L}_{SL} = - \sum_{j=1}^{|\mathcal{X}|} \sum_{i=1}^{n_j} \log p_{\theta}(x_i^j | x_{<i}^j) \quad (12)$$

This objective provides a reliable initialisation by training the model to generate coherent completions based on dialogue context and intermediate representations from dataset demonstrations.

### 3.3.2 Reinforcement Learning with Emotion Reward Signal

Beyond large-scale pretraining and model size, LLMs’ effectiveness also stems from RL-based alignment with human preferences (Ouyang et al., 2022b). Language generation can be framed as a sequential decision-making process, where the model (policy  $\pi$ ) generates tokens (actions  $a$ ) based on context (states  $s$ ) following  $a \sim \pi(\cdot | s)$ , and a reward model, often trained on human preference, guides optimisation. We adopt a similar RL set-up, using user sentiment as a reward to encourage system responses that are both task-effective and emotionally appropriate.

**Episode Formulation** Here, an episode corresponds to the generation of the system response at the  $t$ -th user-system interaction in a dialogue, resulting in the following full sequence:

$$\{\mathcal{H}_t, e_t, d_t, ds_t, db_t, da_t, c_t, u_{\text{sys},t}\} \quad (13)$$

To keep episodes manageable and avoid excessive context length, we define the initial state as:

$$s_t^0 = \{\mathcal{H}_t, e_t, d_t, ds_t, db_t\} \quad (14)$$

Note that superscripts denote the time step in the generation of a single turn, while the subscript  $t$  refers to the dialogue-level time step, which remains fixed during that turn.

Starting from the initial state, each action token  $a_t^l$  is generated from the policy  $\pi_{\theta}(a_t^l | s_t^l)$ . We obtain the reward  $r_t^{l+1}$ , and the state is incrementally updated:

$$s_t^l = \{s_t^l, a_t^{<l}\} \quad (15)$$

At the end of the turn-level generation, we have a list of  $L$  action tokens:

$$a_t^{0:L-1} = \{da_t, c_t, u_{\text{sys},t}\} \quad (16)$$

The LLM-based policy  $\pi_{\theta}$  then generates an action (i.e., text token) according to  $\pi_{\theta}(a_t^l | s_t^l)$ .

### Reward Modelling via Emotion Advantage

We expect the availability of user utterance and emotion prediction from a user simulator (US). Given the system response  $u_{\text{sys},t}$ , the US produces the next user turn and emotion prediction:

$$u_{\text{usr},t+1}, e_{t+1} = \text{US}(u_{\text{sys},t}) \quad (17)$$

To explore diverse behaviours, we sample multiple turn-level trajectories from the same initial state:

$$\tilde{a}_t^{0:L-1} = \{\tilde{da}_t, \tilde{c}_t, \tilde{u}_{\text{sys},t}\} \sim \prod_{l=1}^L \pi(a_t^l | s_t^l) \quad (18)$$

One of the sampled sequences is then randomly selected to continue the conversation with the US. Each sampled response  $\tilde{u}_{\text{sys},t}$  is evaluated by the US to produce an emotion label  $\tilde{e}_{t+1}$ , which is mapped to a valence-based numeric reward with the mapping function  $\mathcal{M}(\cdot)$ : +1 for positive, 0 for neutral, and -1 for negative. Intermediate rewards  $r^{1:L-1} = 0$ , and the final reward is normalised across samples:

$$\tilde{r}_t^L = \frac{\mathcal{M}(\tilde{e}_t) - \mathbb{E}[\mathcal{M}(\tilde{e}_t)]}{\sigma(\mathcal{M}(\tilde{e}_t))} \quad (19)$$

We refer to this signal as the **emotion advantage**, as it captures how much emotionally better (or worse) a generated response is compared to alternatives under the same context. In practice, we drop the set of sampled responses from the same context that lead to the same user emotion (zero variance in  $\mathcal{M}(\tilde{e}_t)$ ).

**User Emotion Optimisation** We aim to maximise the expected return of the policy  $\pi_{\theta}$  across the sampled trajectories with turn-level discount factor  $\gamma_{\text{turn}}$ .

$$J(\theta) = \mathbb{E}_{\pi_{\theta}} \sum_{l=0}^{L-1} \gamma_{\text{turn}}^l r_t^{l+1} \quad (20)$$

Gradients are estimated using the policy gradient theorem and the advantage function  $A_{\pi_{\theta}}(s_t^l, a_t^l)$  (Sutton et al., 1999):

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi_{\theta}} \sum_{l=0}^{L-1} \nabla_{\theta} \log \pi_{\theta}(a_t^l | s_t^l) A_{\pi_{\theta}}(s_t^l, a_t^l) \quad (21)$$

In practice, we apply the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) to approximate the direction of the gradient above with a gradient of the surrogate loss. PPO stabilises training by clipping large policy updates and penalising Kullback–Leibler (KL) divergence between successive policies. For full implementation details, please refer to Appendix A.2.

### 3.3.3 Reinforcement Learning with Multi-turn Reward Signal

While single-turn RL can encourage the system to elicit positive user emotion, effective goal fulfillment requires planning over longer horizons. We therefore extend our training set-up to incorporate multi-turn RL with dialogue-level reward.

**Episode Formulation** An episode corresponds to a full dialogue, consisting of  $T$  user-system turns. At the  $t$ -th step in the dialogue, we derive the state and action with a formatter function  $\mathcal{F}(\cdot)$  from Equations 14 and 16: the dialogue-level state  $\mathcal{S}_t = \mathcal{F}(s_t^0)$  and dialogue-level action  $\mathcal{A}_t = \mathcal{F}(a_t^{0:L-1})$  (see Appendix A.3 for formatter details). This forms a sequence of state-action pairs  $\{(\mathcal{S}_0, \mathcal{A}_0), \dots, (\mathcal{S}_{T-1}, \mathcal{A}_{T-1})\}$  over the course of the interaction.

**Reward Modelling** To guide long-term behaviour, we introduce dialogue-level reward signals based on task success and interaction efficiency. We train two dialogue-level critic networks for our policy  $\pi$  parameterised by  $\phi$  and  $\psi$ : state-action value function  $Q_\phi^\pi(\mathcal{S}_t, \mathcal{A}_t)$  and state value function  $V_\psi^\pi(\mathcal{S}_t)$ . Both critics are based on ALBERT (Lan et al., 2020), with a linear projection on the [CLS] token embedding to produce scalar value estimates. Following the Archer framework (Zhou et al., 2024), they are trained to minimise temporal difference (TD) error and ensure Bellman consistency using online experience  $\mathcal{E} = \{\mathcal{S}_t, \mathcal{A}_t, r_{t+1}, \mathcal{S}_{t+1}\}_{t=0}^{T-1}$  collected during interactions with a user simulator:

$$\mathcal{L}_Q = \mathbb{E}_{s,a,r,s' \sim \mathcal{E}} [(Q_\phi(s, a) - r - \gamma_{\text{dial}} V_\psi(s'))^2] \quad (22)$$

$$\mathcal{L}_V = \mathbb{E}_{s \sim \mathcal{E}} [\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} [(V_\psi(s) - Q_\phi(s, a))^2]] \quad (23)$$

The delayed target models  $Q_\phi^-$  and  $V_\psi^-$  are updated towards their current counterparts with Polyak averaging (Haarnoja et al., 2018). For full critic training details, refer to the work of Zhou et al. (2024) and Appendix A.3.

**Task Completion Optimisation** Given the trained critics, we compute the advantage estimate at each  $t$ -th turn using:

$$A_t = Q_\phi^\pi(\mathcal{S}_t, \mathcal{A}_t) - V_\psi^\pi(\mathcal{S}_t) \quad (24)$$

This value measures how much better a particular action is compared to the expected return of the current state. We then use this advantage signal in the same PPO optimisation framework as described in the set-up for emotion reward (Section 3.3.2), replacing the emotion-based advantage with this task-oriented multi-turn advantage.

**Combined Reward** The single-turn (Equation 19) and multi-turn reward (Equation 24) signals are combined for selected sequences from Equation 18 using a weighted sum with mixing parameter  $\rho$  (set to 0.1 in practice):

$$\mathcal{R}_{\text{combine},t} = \rho A_t + r_t^L \quad (25)$$

**Off-policy Experience Replay** Due to the high computational cost of simulated interactions, we maintain an experience buffer and apply importance sampling (IS) (Precup et al., 2000) to reuse off-policy data. A constant buffer size is enforced using reservoir sampling (Isele and Cosgun, 2018). The IS weight is defined as:

$$\alpha_t = \frac{\pi_{\theta_{\text{new}}}(\mathcal{A}_t|\mathcal{S}_t)}{\pi_{\theta_{\text{old}}}(\mathcal{A}_t|\mathcal{S}_t)} \quad (26)$$

In practice, computing IS weights over long sequences can be numerically unstable due to the sharp trajectory distribution of LLMs. This often leads to high variance in IS weight estimates, even when the behavior and target policies only diverge slightly. To mitigate this, we compute IS weights only over intents and slots of dialogue actions  $da_{t,\text{intent+slot}}$ , and clip them to  $[1 - \eta, 1 + \eta]$  with  $\eta = 0.9$ :

$$\hat{\alpha}_t = \text{clip}\left(\frac{\pi_{\theta_{\text{new}}}(da_{t,\text{intent+slot}}|\mathcal{S}_t)}{\pi_{\theta_{\text{old}}}(da_{t,\text{intent+slot}}|\mathcal{S}_t)}, 1 - \eta, 1 + \eta\right) \quad (27)$$

When deploying experience replay, we multiply the corresponding reward for past experiences (Equation 24 or 25) with IS weight  $\hat{\alpha}_t$ .

## 4 Experiments Set-up

### 4.1 ToD Systems

Based on design considerations discussed in Section 1, we consider the following ToD systems. System configurations are summarised in Table 1.

#### Naive Baseline

**Chitchat-ToD** has a 3B Phi-3.5-mini backbone and is trained to generate response directly, omitting all sub-tasks and the interaction with the database. We aim to investigate if an LLM can memorise the database solely from the corpus.

#### Agentic System

**FnCTOD** (Li et al., 2024) is an agentic system with a prompt-based zero-shot generation process: DST, database query, and response generation. We follow the prompt specified in the original paper and use GPT-4o-mini (2025-04-16).

	Architecture	Optimisation	Dialogue State	Dialogue Action	Emotion	RL Objectives		Training Env	#Param
						Sent	Succ		
Chitchat-ToD	E2E	E2E	Implicit	Implicit	Implicit	-	-	NA	3.82B
FnCTOD (2024)	Agentic	No	Explicit	Implicit	Implicit	-	-	NA	8B(est.)
Modular-base (2022)	Modular	Modular	Explicit	Delexicalised	No	+	-	EmoUS	413M
Modular-emotion (2024a)	Modular	Modular	Explicit	Delexicalised	Explicit	+	+	EmoUS	547M
SimpleLoop (2022)	Modular	Loop	Explicit	Delexicalised	No	+	-	langEmoUS	413M
EmoLoop (2024a)	Modular	Loop	Explicit	Delexicalised	Explicit	+	+	langEmoUS	547M
SimpleLLAMA (2024)	E2E	E2E	Explicit	Delexicalised	Implicit	-	-	NA	7.02B
EmoLLAMA (2024)	E2E	E2E	Explicit	Delexicalised	Explicit	-	-	NA	7.02B
LUSTER-base	E2E	E2E	Explicit	Lexicalised	Implicit	-	-	NA	3.82B
LUSTER-base-emotion	E2E	E2E	Explicit	Lexicalised	Explicit	-	-	NA	3.82B
LUSTER-RL(sent)	E2E	E2E	Explicit	Lexicalised	Explicit	-	+	langEmoUS	3.82B
LUSTER-RL(succ)	E2E	E2E	Explicit	Lexicalised	Explicit	+	-	langEmoUS	3.82B
LUSTER	E2E	E2E	Explicit	Lexicalised	Explicit	+	+	langEmoUS	3.82B

Table 1: System configurations of ToD systems as discussed in Section 1. **Architecture:** end-to-end (*E2E*), *Modular*, or *Agentic*. **Optimisation:** *Modular* optimisation following Figure 2a, *Loop* optimisation following Figure 2b, end-to-end (*E2E*) following Figure 2c,2d or *No* task-specific optimisation. **Dialogue State:** *Explicit* representation or *Implicit* model hidden representations. **Dialogue Action:** *Delexicalised* where no slot values in policy output, *Lexicalised* where slot values are included, or *Implicit* model hidden representations. **Emotion:** *Explicit* policy input/output, *Implicit* consideration from natural language, or *Not* perceived nor expressed by the policy. **RL Objectives:** RL using *Sentiment* or task *Success* reward signal. **Training Environment** for RL training: *NA* when no RL involved, *EmoUS* for semantic interaction, and *langEmoUS* for natural-language interaction. For number of parameters **#Param**, the size of GPT-4o-mini used as FnCTOD backbone is *estimated*.

### Modular Systems with Modular Optimisation

**Modular-base** has SetSUMBT DST (van Niek-erk et al., 2021), DDPT policy (Geishauer et al., 2022), and SEC-BART NLG (Feng et al., 2024a). The policy is trained via RL using *semantic* interaction with an emotional user simulator EmoUS (Lin et al., 2023), receiving a success signal (Figure 2a).

**Modular-emotion** has all components of **Modular-base** with additionally user emotion (Feng et al., 2023) as policy input and system conduct as policy output. The policy also receives turn-level user sentiment as reward.

### Modular Systems with Loop Optimisation

**SimpleLoop** has the same architecture as **Modular-base** but its RL training involves success reward from natural-language interaction (Figure 2a) with an emotional user simulator langEmoUS (Feng et al., 2024a).

**EmoLoop** has the same pipeline set-up as **Modular-emotion** and the same training set-up as **SimpleLoop** with an additional turn-level sentiment reward signal.

### Delexicalised End-to-end Systems

**SimpleLLAMA** (Stricker and Paroubek, 2024) follows the best set-up of SimpleTOD (Hosseini-Asl et al., 2020) to

auto-regressively generate the state, dialogue action, and response. It uses the 7B LLAMA-2-chat (Touvron et al., 2023) as the backbone. It generates delexicalised dialogue actions and responses, and it is trained via SL on the corpus.

**EmoLLAMA** has two more emotion-related sub-tasks than **SimpleLLAMA**: user emotion recognition and system conduct selection.

### LUSTER and its Variants

**LUSTER-base** completes non-affective ToD modelling sub-tasks as listed in Figure 3. It uses a 3B Phi-3.5-mini and generates lexicalised actions and responses.

**LUSTER-base-emotion** has two more affective sub-tasks than **LUSTER-base**: user emotion recognition and system conduct selections.

**LUSTER-RL(sent)** is based on **LUSTER-base-emotion** and incorporates RL with turn-level **sentiment** reward as discussed in Section 3.3.2.

**LUSTER-RL(succ)** is based on **LUSTER-base-emotion** and incorporates RL with dialogue-level **success** reward as discussed in Section 3.3.3.

**LUSTER** is based on **LUSTER-base-emotion** and it incorporates RL using both turn-level sentiment reward and dialogue-level success reward.

## 4.2 Training and Evaluation Environment

All systems or sub-modules are trained on EmoWOZ (an extended version of MultiWOZ). Interactive evaluation is carried out using ConvLab-3 toolkit (Zhu et al., 2023). The natural-language user model consists of T5NLU (Zhu et al., 2023) and langEmoUS (Feng et al., 2024a). T5NLU reports a slot F1 of 86.5%, putting extra emphasis on the clarity of the system’s natural-language responses. LangEmoUS is a state-of-the-art user simulator and reports a macro F1 of 52.1% on emotion prediction. Details about hyperparameters and computing resource can be found in Appendices A.1 and A.5.

We simulate 3000 dialogues per system during evaluation. We report the success rate (judged by the ConvLab-3 evaluator) and the average sentiment (predicted by the user simulator). For success evaluation, the evaluator verifies whether the user’s goal has been met. This includes checking database constraints and ensuring that a booking was made as specified. The sentiment score is assigned for each turn as follows: +1 if the user is classified as “satisfied”, -1 if “dissatisfied” or “abusive”, and 0 otherwise.

## 5 Results and Discussion

### 5.1 Model Performance

	Success (↑)	Sentiment (↑)	Error (↓)
Chitchat-ToD	20.5%	0.368	NA
FnCTOD	32.3%	0.229	NA
Modular-base	31.8%	0.335	5.35%
Modular-emotion	33.0%	†0.345	†5.14%
SimpleLoop	35.2%	0.346	4.74%
EmoLoop	37.2%	† <b>0.380</b>	†3.36%
SimpleLLAMA	33.0%	0.214	5.11%
EmoLLAMA	34.2%	†0.250	†3.68%
LUSTER-base	47.2%	0.298	1.73%
LUSTER-base-emotion	47.5%	†0.310	† <b>1.38%</b>
LUSTER-RL(sent)	†49.8%	†0.345	1.79%
LUSTER-RL(succ)	†50.8%	†0.324	†1.51%
LUSTER	† <b>51.2%</b>	†0.316	1.77%

Table 2: Task **success**, user **sentiment**, and concept **error** rate of all systems. Best scores across all systems are **bolded**. †denotes statistically significant improvement ( $p < 0.05$ ) from the first system within each group.

**Chitchat-style or agentic model lacks the optimisation necessary for goal-oriented task completion.** Chitchat-ToD performs poorly on success rate (20.5%) despite generating positive

sentiment, as it often generates ungrounded content that appears helpful but fails to meet the user’s search constraints. This underscores that smooth, engaging interactions alone are insufficient. FnCTOD, despite comparable success rate to some systems (32.3%), falls short on user sentiment, likely due to its overly informative and less engaging language (see Section 5.2). These findings suggest effective task fulfillment requires both fact-driven database access and emotionally attuned interaction strategies.

**Loop-based optimisation enables more coherent behaviour.** SimpleLoop and EmoLoop, which train policies within the full natural-language interaction loop, outperform their modularly trained counterparts despite having the same architecture. This shows that loop optimisation provides holistic feedback, leading to more coherently optimised behaviours.

**Emotion modelling improves user sentiment across architectures.** Across both modular (e.g., SimpleLoop vs. EmoLoop) and end-to-end (e.g., SimpleLLAMA vs. EmoLLAMA) systems, incorporating user emotion recognition and system conduct prediction consistently improves user sentiment. EmoLoop achieves the highest sentiment score, highlighting the effectiveness of affect-aware modelling for generating emotionally intelligent responses.

**Lexicalised systems enable better grounding and task success.** LUSTER variants significantly outperform delexicalised ones in both success and error rates. Providing full lexical access to slot values improves grounding and overall task effectiveness. This suggests that while symbolic abstractions such as dialogue states and actions remain useful, the enhanced reasoning and representation capacity of LLMs benefits from direct access to lexical content, enabling more accurate and contextually consistent responses.

**RL enhances task success and user sentiment of LUSTER variants’ performance.** Incorporating turn-level sentiment rewards boosts user sentiment, while dialogue-level success signals increase task success to 50.8%. The full model, LUSTER, optimises both objectives, achieving the highest task success (51.2%) while maintaining stable sentiment and concept error. This demonstrates the effectiveness of hierarchical RL in balancing short-term emotional response with long-term goal achievement.

## 5.2 Linguistic Style

	#Turns	#Tokens	Vocab	LogTTR
Chitchat-ToD	17.4	12.9	4,114	0.853
FnCTOD	6.3	29.9	1,301	0.831
Modular-base	16.3	14.0	2,333	0.823
Modular-emotion	16.2	14.2	2,486	0.824
SimpleLoop	15.3	12.8	3,628	0.833
EmoLoop	14.5	11.6	2,256	0.845
SimpleLLAMA	14.6	13.7	2,109	0.834
EmoLLAMA	14.0	12.8	2,035	0.838
LUSTER-base	11.1	11.2	1,841	0.881
LUSTER-base-emotion	11.4	11.3	1,928	0.873
LUSTER-RL(sent)	11.3	11.0	2,043	0.877
LUSTER-RL(succ)	11.6	12.4	2,048	0.870
LUSTER	11.5	12.5	2,026	0.872

Table 3: Average number of user-system exchanges per dialogue (#Turns), average number of tokens per system turn (#Tokens), average number of unique tokens (Vocab), and average log type-token ratio (LogTTR) per dialogue of all systems.

Chitchat-ToD takes the most turns per dialogue (17.4 on average) due to the lack of database grounding for effective user goal fulfillment. In contrast, FnCTOD, completes dialogue on average in 6.3 turns using overly informative responses (nearly 30 tokens per turn). Loop-optimised systems ({Simple,Emo}Loop), compared with their modularly trained counterparts (Modular-{base,emotion}), achieve better performance with fewer turns, highlighting the better efficiency of policies as a result of natural-language optimisation. LUSTER variants, benefiting from lexicalised end-to-end modeling and RL, are the most efficient. They achieve overall best success with an average around 11 turns.

Most systems use around 2,000 unique tokens, reflecting lexical compactness for efficient and goal-directed interactions. SimpleLoop uses over 3,600 tokens due to frequent but unnecessary mentions of phone numbers. Chitchat-ToD, being open-ended and unconstrained, uses the largest vocabulary (4,114). FnCTOD has the smallest (1,301) because of long templated responses.

We also measure the within-dialogue log ratio between the number of unique and total tokens (LogTTR) for a fairer comparison of lexical diversity. LUSTER variants generally achieve high LogTTR ( $> 0.87$ ), indicating greater within-dialogue lexical variation. In contrast, systems like Modular-base exhibit lower diversity despite larger vocabularies than LUSTER.

## 5.3 System Comparison Using LLM Raters

LUSTER variants outperform other models in task success and concept error (Table 2). However, differences among LUSTER variants are small, likely due to limitations of the rule-based evaluator and pretrained user simulator in capturing fine behavioural nuances. To obtain more sensitive and human-aligned assessment, we employ proprietary LLMs as judges. Each system interacts with the user simulator under the same user goal. Their dialogues are presented in random order to GPT-4o-mini (2025-04-16) and Deepseek-R1 (2025-05-28), tasked with identifying the better system (see Appendix A.6.1).

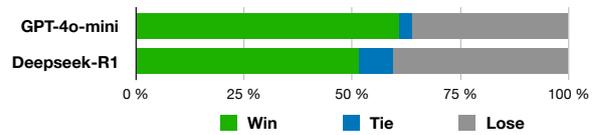


Figure 4: LUSTER vs. LUSTER-base-emotion.

Figure 4 shows the win-rate comparison between LUSTER and its non-RL counterpart, LUSTER-base-emotion for 500 dialogues. Both LLM raters exhibit a clear preference for LUSTER: GPT-4o-mini attributes a strong majority of wins to LUSTER, while Deepseek-R1 also favors LUSTER, albeit with a more moderate margin. These results confirm that RL contributes to nuanced yet perceptible improvements in conversational quality. Appendix A.6.2 exemplifies how LLMs identify behavioural nuances.

## 6 Conclusion

In this work, we systematically investigate the design dimensions of ToD systems, spanning architectural choices, optimisation strategies, representation, and emotion modeling. Through experimentation across modular, end-to-end, and agentic systems, we find that no single design axis alone guarantees optimal performance. Instead, efficient task success, user satisfaction, and semantic accuracy emerge from their thoughtful combination. Our proposed LUSTER system demonstrates this by integrating lexicalised generation, emotion-aware modeling, and hierarchical reinforcement learning, achieving a significantly higher task success rate and overall balanced performance. These findings underscore the importance of holistic, end-to-end, yet structured design in building conversational agents capable of effective and emotionally intelligent task completion.

## References

- Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, et al. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). Technical Report MSR-TR-2024-12, Microsoft.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Federico Errica, Davide Sanvito, Giuseppe Siracusa, and Roberto Bifulco. 2025. [What did I do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558, Albuquerque, New Mexico. Association for Computational Linguistics.
- Shutong Feng, Hsien-chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Matthias Ruppik, Renato Vukovic, and Milica Gasic. 2024a. [Infusing emotions into task-oriented dialogue systems: Understanding, management, and generation](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 699–717, Kyoto, Japan. Association for Computational Linguistics.
- Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. [EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.
- Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gasic. 2023. [From chatter to matter: Addressing critical steps of emotion recognition learning in task-oriented dialogue](#). In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 85–103, Prague, Czechia. Association for Computational Linguistics.
- Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. 2024b. [Affect recognition in conversations using large language models](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 259–273, Kyoto, Japan. Association for Computational Linguistics.
- Artur S. d'Avila Garcez, Dov M. Gabbay, and Krysia B. Broda. 2002. *Neural-Symbolic Learning System: Foundations and Applications*. Springer-Verlag, Berlin, Heidelberg.
- M. Gašić, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young. 2013. [On-line policy optimisation of bayesian spoken dialogue systems via human interaction](#). In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371.
- Christian Geishauser, Carel van Niekerk, Hsien-chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gašić. 2022. [Dynamic dialogue policy for continual reinforcement learning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 266–284, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. [Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor](#). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1856–1865. PMLR.
- Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk,

- and Milica Gasic. 2023. [ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Vojtěch Hudeček and Ondrej Dusek. 2023. [Are large language models all you need for task-oriented dialogue?](#) In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.
- David Isele and Akansel Cosgun. 2018. Selective experience replay for lifelong learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18*. AAAI Press.
- Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2022. [GPT-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems](#). In *International Conference on Learning Representations*.
- Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.
- Florian Kreyszig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. [Neural user simulation for corpus-based policy optimisation of spoken dialogue systems](#). In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 60–69, Melbourne, Australia. Association for Computational Linguistics.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [Albert: A lite bert for self-supervised learning of language representations](#). In *International Conference on Learning Representations*.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Zekun Li, Zhiyu Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Dong, Adithya Sagar, Xifeng Yan, and Paul Crook. 2024. [Large language models as zero-shot dialogue state tracker through function calling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704, Bangkok, Thailand. Association for Computational Linguistics.
- Hsien-Chin Lin, Shutong Feng, Christian Geishauer, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasić. 2023. [Emous: Simulating user emotions in task-oriented dialogues](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2526–2531, New York, NY, USA. Association for Computing Machinery.
- Hsien-Chin Lin, Chia-Hao Shen, Benjamin Matthias Ruppik, Carel van Niekerk, Michael Heck, Nurul Lubis, Renato Vukovic, Shutong Feng, and Milica Gasic. 2024. [Text as parameter: interactive prompt optimisation for large language models](#).

- Nurul Lubis, Christian Geishauer, Michael Heck, Hsien-chin Lin, Marco Moresi, Carel van Niekerk, and Milica Gasic. 2020. [LAVA: Latent action spaces via variational auto-encoding for dialogue policy optimization](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 465–479, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Carel van Niekerk, Andrey Malinin, Christian Geishauer, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. [Uncertainty measures in neural belief tracking and the effects on dialogue policy performance](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. 2022a. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, et al. 2022b. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. [Soloist: Building task bots at scale with transfer learning and machine teaching](#). *Transactions of the Association for Computational Linguistics*, 9:807–824.
- Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.
- Doina Precup, Richard S. Sutton, and Satinder P. Singh. 2000. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00*, page 759–766, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- David Premack and Guy Woodruff. 1978. [Does the chimpanzee have a theory of mind?](#) *Behavioral and Brain Sciences*, 1(4):515–526.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. [Agenda-based user simulation for bootstrapping a POMDP dialogue system](#). In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York. Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#).
- Armand Stricker and Patrick Paroubek. 2024. [A Unified Approach to Emotion Detection and Task-Oriented Dialogue Modeling](#). In *IWSDS*, Sapporo (Japan), Japan.
- Guangzhi Sun, Shutong Feng, Dongcheng Jiang, Chao Zhang, Milica Gasic, and Phil Woodland. 2024. [Speech-based slot filling using large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6351–6362, Bangkok, Thailand. Association for Computational Linguistics.
- Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'99*, page 1057–1063, Cambridge, MA, USA. MIT Press.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Renato Vukovic, David Arps, Carel van Niekerk, Benjamin Matthias Ruppik, Hsien-chin

- Lin, Michael Heck, and Milica Gasic. 2024. [Dialogue ontology relation extraction via constrained chain-of-thought decoding](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 370–384, Kyoto, Japan. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. [A network-based end-to-end trainable task-oriented dialogue system](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.
- Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. [Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–677, Vancouver, Canada. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. [Pomdp-based statistical spoken dialog systems: A review](#). *Proceedings of the IEEE*, 101(5):1160–1179.
- Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 2006. *The HTK Book Version 3.4*. Cambridge University Press.
- Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. [ArCHer: Training language model agents via hierarchical multi-turn RL](#). In *Forty-first International Conference on Machine Learning*.
- Qi Zhu, Christian Geisshauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Shutong Feng, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gasic, and Minlie Huang. 2023. [ConvLab-3: A flexible dialogue system toolkit based on a unified data format](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 106–123, Singapore. Association for Computational Linguistics.

## A Supplementary Information

### A.1 Computing Resources

All modular systems were trained and deployed using a single NVIDIA RTX 2080 (8GB) graphics processing unit (GPU). The end-to-end systems SimpleLLAMA and EmoLLAMA were trained and evaluated on an NVIDIA A100 (80GB) GPU. LUSTER variants were trained on a single NVIDIA A100 (40GB) GPU, with inference requiring at least an NVIDIA RTX 6000 (24GB) GPU. While training LUSTER demands more computational resources than modular systems, it remains more resource-efficient than LLAMA-based end-to-end models.

### A.2 Proximal Policy Optimisation for Single-turn RL

The advantage function  $A_\pi(s, a)$  denotes the difference between the action-value function  $Q_\pi(s, a)$ , which is the expected return when taking action  $a$  in state  $s$  and from then on following the policy  $\pi$  and the state value function  $V_\pi(s)$ , which is the expected return of state  $s$  for policy  $\pi$ .

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s) \quad (\text{A.1})$$

In the context of single-turn response generation with an episode length of  $L$  (Equation 14) for the  $t$ -th interaction of a dialogue with a policy parameterised with  $\theta$ ,

$$Q_{\pi_\theta}(s, a) = \mathbb{E}_{\pi_\theta} \sum_{k=0}^{L-l-1} [\gamma^k r_t^{k+1} | s = s_t^l, a = a_t^l] \quad (\text{A.2})$$

$$V_{\pi_\theta}(s) = \mathbb{E}_{\pi_\theta} \sum_{k=0}^{L-l-1} [\gamma^k r_t^{k+1} | s = s_t^l] \quad (\text{A.3})$$

$$\hat{A}_t^l = Q_{\pi_\theta}(s, a) - V_{\pi_\theta}(s) \quad (\text{A.4})$$

The surrogate loss consisting of the clipping term and the KL-penalty term is given by:

$$\begin{aligned} \mathcal{L}_{\text{PPO}} = \mathbb{E}_t \left[ \min \left( \frac{\pi_\theta(a_t^l | s_t^l)}{\pi_{\theta_{\text{old}}}(a_t^l | s_t^l)} \hat{A}_t^l, \text{clip} \left( \frac{\pi_\theta(a_t^l | s_t^l)}{\pi_{\theta_{\text{old}}}(a_t^l | s_t^l)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t^l \right) \right. \\ \left. - \beta \cdot \sum_a \pi_{\theta_{\text{old}}}(a | s_t^l) \log \left( \frac{\pi_\theta(a | s_t^l)}{\pi_{\theta_{\text{old}}}(a | s_t^l)} \right) \right], \end{aligned} \quad (\text{A.5})$$

where  $\text{clip}(x, a, b)$  ensures that  $x$  is within the interval  $[a, b]$ . The gradient of the surrogate loss (Eq A.5) approximates the direction of the policy gradient (Eqn 21).

To estimate advantage function  $\hat{A}_t$  needed for the surrogate loss, generalised advantage estimation is used:

$$\hat{A}_t^l = \sum_{j=0}^{L-j-1} (\gamma_{\text{turn}} \lambda)^j \delta_t^{l+j} \quad \text{where} \quad \delta_t^l = r_t^l + \gamma_{\text{turn}} V_\pi(s_{t+1}^l) - V_{\pi_\theta}(s_t^l) \quad (\text{A.6})$$

Parameter  $\lambda \in [0, 1]$  trades off bias (lower  $\lambda$ ) and variance (higher  $\lambda$ ).  $V_{\pi_\theta}$  is initialised with the reward model  $R_\phi$  for complete sequences, and parameters  $\phi$  are updated in such a way so that  $V_{\pi_\theta}$  estimates observed returns for partial sequences. In this way, the critic parameterised with  $\phi$  influences the actor parameterised with  $\theta$ . We follow default hyperparameters in the PPO trainer of TRL library, which follows [Ouyang et al. \(2022a\)](#).

PPO is an online learning algorithm, which means it needs to adopt exploration during the process of optimisation. That is achieved by applying a temperature parameter to the softmax layer of policy  $\pi_\theta$  to induce more varied responses.

In practice, the critic is modelled by a single-layer value head attached to the output of the language model. It projects the logits of the token into a scalar. The input for calculating  $V_{\pi_\theta}(s)$  is the logit of the last token in  $s$ , and that for calculating  $Q_{\pi_\theta}(s, a)$  is the logit of  $a$ , since the action  $a$  is the next token generated.

### A.3 Dialogue-level Critics

#### A.3.1 Architecture

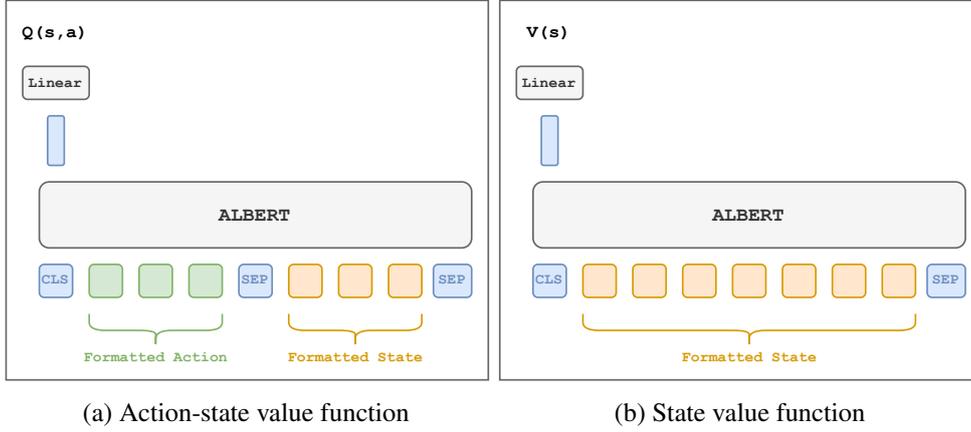


Figure A.1: Dialogue-level Critics

Figure A.1 illustrates the modelling of dialogue critics. We format the dialogue-level state  $\mathcal{S}_t$  and action  $\mathcal{A}_t$  from the generated sequence  $s_t^0$  and  $a_t^{1:L-1}$  to obtain a compact input to the critic.

**Textual State Representation** The input to the state value function is in the form of the following:

$\langle domain \rangle \langle slot \rangle \langle \text{"inactive"} | \text{"informed"} | \text{"unknown"} \rangle ; \dots [\text{SEP}] \langle N \langle domain \rangle \text{found} \rangle | \text{"no entity found"} | \text{"database not available"} [\text{SEP}] \text{"system :"} u_{\text{sys}, t-1} [\text{SEP}] \text{"user :"} u_{\text{usr}, t}$

The first segment contains the system’s belief about the user’s goal. Instead of actual informed value, we limit the status of each slot to be one of the three: *informed*, *unknown* if not informed, or *inactive* if the  $\langle domain \rangle$  is not the active domain determined by the system  $\langle domain \rangle \neq d_t$ .

The second segment contains a summary of the database query status. The status informs the number of entities returned  $N$  for  $N > 0$  or informs “no entity found” if  $N = 0$ . There is a special case when the active domain is “general” for small-talk and greetings. Under such circumstances, the status informs “database not available”.

To account briefly for the dialogue history, we also include the previous system and user turns  $u_{\text{sys}, t-1}$  and  $u_{\text{usr}, t}$ .

**Textual Action Representation** The formatted action for the state-action value function is in the form of the following. Values are excluded.

$\langle domain \rangle \langle intent \rangle \langle slot \rangle ; \dots$

#### A.3.2 Training

Following the objective functions defined in Equation 22 and 23, for each critic update step:

$$\phi \leftarrow \phi - \nabla J_\phi(Q) \quad (\text{A.7})$$

$$\psi \leftarrow \psi - \nabla J_\psi(V) \quad (\text{A.8})$$

Parameters of target critics are then updated via Polyak averaging (Haarnoja et al., 2018):

$$\bar{\phi} \leftarrow (1 - \tau)\bar{\phi} + \tau\phi \quad (\text{A.9})$$

$$\bar{\psi} \leftarrow (1 - \tau)\bar{\psi} + \tau\psi \quad (\text{A.10})$$

## A.4 Pseudo Code

---

### Algorithm 1 LUSTER Training

---

```

1: Initialise parameters  $\theta, \phi, \psi, \bar{\phi}, \bar{\psi}$ 
2: ## Actor Pretraining using EmoWOZ
3: for each actor step do
4:    $\theta \leftarrow \theta - \nabla \mathcal{L}_{SL}$  ▷ Equation 12
5: end for
6: ## Critic Pretraining using EmoWOZ, not applicable when using emotion advantage only.
7: for each critic step do
8:   ## Update utterance-level Q and V functions by target function bootstrapping.
9:    $\phi \leftarrow \phi - \nabla \mathcal{L}_Q$  ▷ Equation 22
10:   $\psi \leftarrow \psi - \nabla \mathcal{L}_V$  ▷ Equation 23
11:  ## Update target Q and V functions via Polyak averaging.
12:   $\bar{\phi} \leftarrow (1 - \tau)\bar{\phi} + \tau\phi$  ▷ Equation A.9
13:   $\bar{\psi} \leftarrow (1 - \tau)\bar{\psi} + \tau\psi$  ▷ Equation A.10
14: end for
15: ## Initialise and fill experience replay buffer  $\mathcal{D}$ 
16: for each dialogue while buffer_size not reached do
17:   for each turn do
18:    Execute  $a_t^{0:L-1} \sim \pi_\theta(\cdot | s_t^0)$ , obtain the next state  $s_{t+1}^0$  and reward  $\{r_t^L, A_t, \text{ or } \mathcal{R}_{\text{combine}, t}\}$ , add to buffer  $\mathcal{D}$ .
19:   end for
20: end for
21: for each iteration using buffer  $\mathcal{D}$  do
22:   Repeat line 17-21 until  $\mathcal{D}$  is full.
23:   Repeat line 7-14 for critic update.
24:   for each off-policy experience do
25:    Replay experience and sale reward with importance sampling weight. ▷ Equation 27
26:   end for
27:   ## Update token-level actor with utterance-level critic.
28:   for each actor step do
29:     $\theta \leftarrow \theta - \nabla \mathcal{L}_{PPO}$  ▷ Equation 21, approximated with A.5
30:   end for
31:   Discard oldest experiences based on specified Training Interval to make room for new experiences
32: end for

```

---

## A.5 Training Configurations

The predefined set for token selection in the following equations are:

In Equation 2,  $E = \{\text{neutral, fearful, dissatisfied, apologetic, abusive, excited, satisfied}\}$ , defined in the EmoWOZ (Feng et al., 2022).

In Equation 3,  $D = \{\text{general, restaurant, hotel, train, attraction, taxi, police, hospital}\}$ , which includes the seven EmoWOZ domains plus an additional “general” domain to handle small-talk and greetings.

In Equation 10,  $C = \{\text{neutral, compassionate, apologetic, enthusiastic, appreciative}\}$  as defined in EmoWOZ 2.0 (Feng et al., 2024a).

Parameter	Value
Learning Rate	3e-5
Training Epochs	10
Batch Size	32
LoRA Rank	32
LoRA Alpha	32
LoRA Dropout	0.1
LoRA Bias	None
LoRA Target Modules	o_proj, qkv_proj

Table A.1: Supervised Training Hyperparameters

Parameter	Value
Critic Learning Rate	5e-5
Critic Batch Size	32
Critic Epochs	5
Critic Gradient Norm Clipping	40
Actor Learning Rate	1e-7
Actor Batch Size	16
Actor Sample Size (Equation 18)	6
Actor Sample Temperature (dialogue action $da$ )	0.5
Actor Sample Temperature (response $u_{\text{sys}}$ )	0.9
Actor History Length	5
Training Interval	10 Dialogues
Experience Buffer Size	50 Dialogues
Reservoir Sampling	True
User Utterance and Emotion Sample Size (Equation 18)	6
Dialogue-level Discount Factor $\gamma_{\text{dial}}$ (Equation 22)	0.99
Reward Combination Weight $\rho$ (Equation 25)	0.1
Importance Sampling Weight Clipping Range $\eta$ (Equation 27)	0.9
Turn-level Discount Factor $\gamma_{\text{turn}}$ (Equation A.6)	1.0
Decay Rate $\lambda$	0.99
Dialogue-level Reward (Equation (22))	$r_t = -1$ for $0 < t < T$ , $r_T = +80$ if success else -40
Polyak Averaging Weight $\tau$ (Equation A.9, A.10)	0.01

Table A.2: Reinforcement Learning Configurations

## A.6 System Comparison

### A.6.1 Prompt Template

*There are two systems system\_A and system\_B. The user talks to each of them to complete the same information retrieval goal. You are now act as a judge that determines which system has better overall performance.*

*The dialogue between the user and system\_A in json list form is: {system\_a\_dialogue}  
The dialogue between the user and system\_B in json list form is: {system\_b\_dialogue}*

*Now, output your judgement together with explanation in the following format:*

```
{
  "judgement": "which system is better, system_A or system_B? If you really cannot decide, you
can output 'tie'.",
  "explanation": "Why one system is better. Your analysis should be based on the success of the
dialogue, the informativeness of the system responses, and the system emotional performance
to enhance user experience. You can also consider the length of the dialogue, the diversity of
language, and whether the system provided relevant information to the user."
}
```

### A.6.2 Case Study

The simulated user goal is described in the JSON format below where the user’s goal is to retrieve **requestable** information based on **informable** constraints:

```
{
  "restaurant": {
    "informable": {
      "area": "centre",
      "pricerange": "moderate"
    },
    "requestable": {"postcode": "?"}
```

```

    },
    "train": {
      "infomable": {
        "arriveBy": "20:45",
        "day": "thursday",
        "departure": "broxbourne",
        "destination": "cambridge"
      },
      "requestable": {"trainID": "?"}
    }
  }
}

```

The conversation between the simulated user and LUSTER-base-emotion system is:

**User** : I'm looking for a restaurant in the centre of town.

**LUSTER-base-emotion** : there are 69 restaurants in that area. what type of food would you like?

**User** : I would like a moderately priced one please.

**LUSTER-base-emotion** : there is 21 of those. what type of cuisine would you like?

**User** : Could I have the postcode of one of them?

**LUSTER-base-emotion** : [sure, the postcode is cb21db.](#)

**User** : Thank you. I also need a train. I need to leave for cambridge on Thursday and arrive by 20:45.

**LUSTER-base-emotion** : where will you be departing from?

**User** : I will be departing from broxbourne going to cambridge.

**LUSTER-base-emotion** : tr0393 leaves at 05:32 and arrives at 06:32. [would you like to book that for 1 person?](#)

**User** : thank you. bye

**LUSTER-base-emotion** : [thank you for using our service.](#)

Under the same goal, the conversation between the simulated user and LUSTER system is:

**User** : I am looking for information on a place to eat that is in the centre.

**LUSTER** : there are 69 restaurants in that area. what type of food would you like?

**User** : Any kind of food in the moderate price range would be fine.

**LUSTER** : there is 21 of those. what type of cuisine would you like?

**User** : I don't have a preference on cuisine, I just need the postcode.

**LUSTER** : [pizza express has the postcode cb21db.](#)

**User** : Great! Thank you. I also need a train on Thursday that is arriving by 20:45.

**LUSTER** : where will you be departing from?

**User** : I will be departing from broxbourne going to cambridge.

**LUSTER** : tr0393 leaves at 05:32 and arrives at 06:32. [would you like me to book that for you? how many tickets do you need?](#)

**User** : thank you. bye

**LUSTER** : [thank you for using our service. goodbye.](#)

Both systems fulfill the user goal by correctly providing requestable information: postcode=cb21db and trainID=tr0393. Convlab-3 evaluator marks both conversation as successful, and the user simulator, langEmoUS, gives the same sentiment rating in each turn. Despite this, GPT-4o-mini prefers LUSTER over LUSTER-base based on more nuanced system behaviours as highlighted in blue colour above with the following explanation:

LUSTER outperforms LUSTER-base in several key areas.

Firstly, **LUSTER provides a specific restaurant name** (Pizza Express) along with the postcode, which adds a layer of useful information for the user that LUSTER-base lacks, as it only provides the postcode without a specific restaurant name associated with it.

Secondly, LUSTER uses slightly more user-friendly language, such as **'how many tickets do you need?'** which invites the user into the conversation and indicates a more engaging approach.

In terms of emotional performance, LUSTER ends the conversation with a more polite **'goodbye'** which may enhance user experience by making it feel more personal. Overall, LUSTER demonstrates better informativeness, engagement, and emotional connection with the user, leading to a more successful interaction.

## Chapter 10

# Conclusion

In this chapter, we provide a summary of the key findings and contributions of this thesis. We also discuss the limitations of current works and point out directions for future work. The objective of the research presented here was to develop an emotionally intelligent task-oriented dialogue (ToD) system.

### 10.1 Summary of Key Findings

This thesis investigates how emotion can be effectively integrated into ToD systems to enhance both task efficiency and user satisfaction. Across multiple studies, it has advanced the field from foundational resource construction to methodological innovation and system-level integration.

To start with, we construct EmoWOZ, the first large-scale corpus specifically designed for emotion recognition in ToDs (Chapter 5). EmoWOZ addressed the scarcity of suitable datasets by introducing a dedicated taxonomy tailored to the subtle affective expressions inherent in goal-oriented interactions. The resulting resource has provided a benchmark for evaluating emotion recognition in conversation (ERC) models and has facilitated the development of supervised and transfer learning approaches for emotion-aware dialogue modelling.

Following the construction of EmoWOZ, we propose novel learning frameworks for emotion recognition in ToDs, addressing challenges in data imbalance, feature representation, and loss optimisation (Chapter 6). Our framework can be applied to existing ERC models, and the resulting models demonstrate significantly improved performance. In addition, we also carry out a comparative study with large language models (LLMs) and reveal their complementary strengths: while specialised supervised models excelled in accuracy and efficiency, LLMs exhibit notable flexibility, adaptability to unseen domains, and robustness to noisy input conditions (Chapter 7).

Finally, this thesis made the first practical attempt to embed emotion understanding, management, and expression throughout the ToD pipeline (Chapter 8). Through modular and end-to-end implementations, it demonstrated that incorporating emotional intelligence can yield measurable improvements in both objective task success and subjective user experience. We further carry out a comprehensive exploration of the optimal combination of design considerations for building ToD systems. We build an end-to-end and lexicalised system and optimise it via reinforcement learning to balance functional and affective goals. Evaluating the system with the corpus, user simulator, and real humans confirms that emotionally intelligent dialogue agents can be both effective and emotionally attuned.

These contributions establish an integrated framework for emotionally aware ToD systems as described in Chapter 9. Experimental results demonstrate that incorporating emotional intelligence can lead to measurable improvements in both task performance and user emotional experience, suggesting that emotionally intelligent behaviour plays an important role in building robust, user-centred ToD interaction.

## 10.2 Limitations

While this thesis has made substantial progress toward developing emotionally intelligent ToD systems, several limitations remain that should be acknowledged.

A first limitation lies in the **domain and contextual scope** of the research as well as the lack of **multimodal or multilingual considerations**. The emotion taxonomy and models proposed were primarily developed and validated within the domains represented in the EmoWOZ corpus, which focuses on English text-based, goal-oriented interactions. Although these settings capture a diverse range of emotional dynamics, they still reflect a restricted subset of human communication. Real-world applications often involve multimodal signals such as vocal prosody, facial expressions, gestures, or physiological cues, which convey emotion in ways not captured by text alone. As a result, while the proposed models perform effectively in textual environments, their transferability to multimodality and expanded tasks remains to be fully established.

The **evaluation methodology** employed in this work also presents constraints. Existing quantitative metrics for ToD systems, such as task success rate and dialogue length, are well-suited to assessing functional performance but insufficient to capture affective quality. The emotional appropriateness of system responses is approximated with the sentiment of the simulated user, which is inherently noisy. The actual perceived empathy by the user and the user's long-term emotional engagement remain difficult to measure objectively. While user studies with human participants partially address this gap, they are often limited in scale and diversity. Developing richer, multidimensional evaluation frameworks that combine computational, behavioural, and psychological perspectives would allow for a more holistic assessment of emotionally intelligent systems.

From a practical standpoint, **computational efficiency and scalability** also pose limitations. While large language model (LLM)-based emotion recognisers demonstrated strong adaptability and robustness in low-resource settings, their high computational cost and latency make them less viable for real-time deployment in resource-constrained environments. Conversely, the lightweight supervised models developed in this work, though efficient, rely on domain-specific data and may lack generalisation beyond their training contexts. Striking a balance between model complexity, responsiveness, and cross-domain adaptability remains an open engineering and research challenge.

Another important limitation concerns the **conceptual scope of emotional intelligence** as implemented in this thesis. The dialogue systems developed in this work are designed to *behave emotionally*, meaning that they can recognise, reason about, and express emotions during interaction. However, they do not possess real emotions or internal affective experiences. Instead, emotions are represented as computational signals that help guide dialogue decisions and response generation. While this behavioural approach is suitable for current task-oriented applications, it raises the broader question of whether future socially intelligent agents should go beyond emotional simulation and incorporate persistent internal affective states.

Finally, this research does not fully resolve the **ethical and social implications** of emotionally intelligent systems. By endowing dialogue agents with the ability to perceive and respond to human emotions, new ethical questions emerge regarding privacy, consent, emotional manipulation, and authenticity. A system capable of detecting a user's emotional vulnerability, for example, could inadvertently exploit it in persuasion or recommendation contexts. These concerns underscore the necessity of embedding ethical principles, such as transparency, accountability, and user autonomy, into the design and governance of future emotionally intelligent ToD systems.

## 10.3 Recommendations for Future Research

Building on these limitations, several promising directions emerge for future research on emotionally intelligent ToD systems.

A natural progression involves the integration of **multimodal and multilingual emotional intelligence**. Future systems should move beyond purely textual interaction to incorporate multiple

modalities, such as speech, facial expressions, gaze, and physiological feedback, while also extending emotional, thereby capturing the full richness of human affective communication. Incorporating multimodal signals can significantly improve emotion understanding and help resolve ambiguities that arise from text-only cues. Multilingual capability, on the other hand, allows emotionally intelligent dialogue systems to generalise across linguistic boundaries, accounting for cultural variations in affect expression and interpretation. Advancing toward this goal will require the creation of unified, cross-lingual and cross-modal datasets and architectures capable of robust multimodal fusion.

Another key direction lies in **continual learning**. Current models, though effective in predefined contexts, still rely on static datasets and fixed ontologies. Future emotionally intelligent systems should be capable of learning continuously from ongoing interactions while safeguarding against catastrophic forgetting. Moreover, cross-domain transfer, where emotion models trained in one task or cultural setting adapt effectively to another, remains an underexplored yet crucial challenge. Methods such as meta-learning, parameter-efficient fine-tuning, and adaptive representation learning could potentially play central roles in enabling such generalisation.

In terms of decision-making, while the thesis introduced multi-objective optimisation that balances task success and emotional appropriateness, future work could develop more **granular emotion-grounded reward functions** that account for long-term user satisfaction, empathy consistency, and emotional well-being. Incorporating human-in-the-loop reinforcement learning signals could further refine such models, allowing systems to adapt not just to functional goals but also to dynamic affective states over time.

The advancement of **evaluation methodologies** is equally important. Future studies should move toward user-centric, longitudinal assessments that measure not only immediate satisfaction but also trust, rapport, and emotional engagement over extended interactions. This could involve combining quantitative dialogue metrics with qualitative feedback, psychometric evaluations, and behavioural indicators. Evaluating systems in realistic deployment settings, such as customer service, healthcare, or education, would provide a more authentic understanding of their strengths and weaknesses.

Beyond improved evaluation, future progress in emotionally intelligent dialogue systems will also depend on addressing questions of **ethics**. Because such systems operate at the intersection of technology and psychology, their influence on users' affective states and well-being can be significant. Ethical design, therefore, requires explicit consideration of privacy, informed consent, and the responsible handling of affective cues, particularly when systems infer or respond to sensitive affective cues. Establishing guidelines that specify how systems express emotion, treat users equitably, and preserve users' sense of control will be crucial for fostering trust and accountability.

Finally, the pursuit of **explainable affective reasoning** is equally important. As affect-aware dialogue models grow in complexity, users and developers must be able to understand the basis for emotional interpretations and decisions. Future research should explore interpretable affective models capable of articulating how emotional states are inferred from multimodal inputs and how these states shape system behaviour. Embedding such transparency within affective reasoning not only supports ethical accountability but also facilitates more trustworthy and emotionally aligned interaction between humans and dialogue systems.



# Bibliography

- Acosta, Jaime (June 2009). "Using Emotion to Gain Rapport in a Spoken Dialog System". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*. Ed. by Ulrich Germann, Chirag Shah, Svetlana Stoyanchev, Carolyn Penstein Rosé, and Anoop Sarkar. Boulder, Colorado: Association for Computational Linguistics, pp. 49–54. URL: <https://aclanthology.org/N09-3009/>.
- Allen, James (1995). *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.
- Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives (2007). "DBpedia: a nucleus for a web of open data". In: *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Busan, Korea: Springer-Verlag, 722–735. ISBN: 3540762973.
- Austin, John Langshaw (1962). *How to do things with words*. William James Lectures. Oxford University Press. URL: [http://scholar.google.de/scholar.bib?q=info:xI2JvixH8\\_QJ:scholar.google.com/&output=citation&hl=de&as\\_sdt=0,5&ct=citation&cd=1](http://scholar.google.de/scholar.bib?q=info:xI2JvixH8_QJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=1).
- Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer normalization". In: *arXiv preprint arXiv:1607.06450*.
- Bahdanau, Dzmitry (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.
- Banse, Rainer and Klaus Scherer (Mar. 1996). "Acoustic Profiles in Vocal Emotion Expression". In: *Journal of personality and social psychology* 70, pp. 614–36. DOI: 10.1037/0022-3514.70.3.614.
- Beeching, Edward, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf (2023). *Open LLM Leaderboard*. [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard).
- Bickmore, Timothy W and Rosalind W Picard (2005). "Establishing and maintaining long-term human-computer relationships". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 12.2, pp. 293–327.
- Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (July 2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: <https://aclanthology.org/2020.acl-main.485/>.
- Blum, Avrim and Tom Mitchell (1998). "Combining labeled and unlabeled data with co-training". In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT' 98. Madison, Wisconsin, USA: Association for Computing Machinery, 92–100. ISBN: 1581130570. DOI: 10.1145/279943.279962. URL: <https://doi.org/10.1145/279943.279962>.
- Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 4356–4364. ISBN: 9781510838819.
- Bottou, Léon (2010). "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics*. Springer. Paris, France, pp. 177–186.

- Bradley, Margaret M, Mark K Greenwald, Margaret C Petry, and Peter J Lang (1992). "Remembering pictures: pleasure and arousal in memory." In: *Journal of experimental psychology: Learning, Memory, and Cognition* 18.2, p. 379.
- Bradley, Margaret M and Peter J Lang (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the center for research in psychophysiology ...
- Bradley, Ralph Allan and Milton E Terry (1952). "Rank analysis of incomplete block designs: I. the method of paired comparisons". In: *Biometrika* 39.3/4, pp. 324–345.
- Broekens, Joost and Willem-Paul Brinkman (2013). "AffectButton: A method for reliable and valid affective self-report". In: *International Journal of Human-Computer Studies* 71.6, pp. 641–667.
- Brown, Penelope and Stephen C Levinson (1987). *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.
- Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Vancouver, Canada: Curran Associates, Inc., pp. 1877–1901. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf).
- Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić (Oct. 2018). "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 5016–5026. DOI: 10.18653/v1/D18-1547. URL: <https://aclanthology.org/D18-1547/>.
- Buechel, Sven and Udo Hahn (Apr. 2017). "EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 578–585. URL: <https://aclanthology.org/E17-2092/>.
- Bunt, Harry (2011). "The semantics of dialogue acts". In: *Proceedings of the Ninth International Conference on Computational Semantics*. IWCS '11. Oxford, United Kingdom: Association for Computational Linguistics, 1–13.
- Bunt, Harry, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot (May 2020). "The ISO Standard for Dialogue Act Annotation, Second Edition". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 549–558. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.69/>.
- Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. (2005). "A database of german emotional speech." In: *Interspeech*. Vol. 5. Lisbon, Portugal, pp. 1517–1520.
- Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan (2008). "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42.4, pp. 335–359.
- Cannon, Walter B. (1927). "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory". In: *The American Journal of Psychology* 39.1/4, pp. 106–124. ISSN: 00029556. URL: <http://www.jstor.org/stable/1415404> (visited on 08/25/2025).
- Cao, Houwei, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma (2014). "Crema-d: Crowd-sourced emotional multimodal actors dataset". In: *IEEE transactions on affective computing* 5.4, pp. 377–390.
- Caruana, Rich (1997). "Multitask learning". In: *Machine learning* 28.1, pp. 41–75.
- Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (Oct. 2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and

- Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: <https://aclanthology.org/D14-1179/>.
- Chowdhery, Aakanksha et al. (2023). “Palm: Scaling language modeling with pathways”. In: *Journal of Machine Learning Research* 24.240, pp. 1–113.
- Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017). “Deep reinforcement learning from human preferences”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 4302–4310. ISBN: 9781510860964.
- Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio (2014). “Empirical evaluation of gated recurrent neural networks on sequence modeling”. In: *arXiv preprint arXiv:1412.3555*.
- Clark, Herbert H. and Susan E. Brennan (1991). “Grounding in communication.” In: *Perspectives on socially shared cognition*. Washington, DC, US: American Psychological Association, pp. 127–149. ISBN: 1-55798-121-3 (Hardcover). DOI: 10.1037/10096-006. URL: <https://doi.org/10.1037/10096-006>.
- Cohen, Jacob (1960). “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1, pp. 37–46.
- Cohen, Philip R and Hector J Levesque (1990). “Intention is choice with commitment”. In: *Artificial intelligence* 42.2-3, pp. 213–261.
- Colby, Kenneth Mark, Franklin Dennis Hilf, Sylvia Weber, and Helena C Kraemer (1972). “Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes”. In: *Artificial Intelligence* 3, pp. 199–221. ISSN: 0004-3702. DOI: [https://doi.org/10.1016/0004-3702\(72\)90049-5](https://doi.org/10.1016/0004-3702(72)90049-5). URL: <https://www.sciencedirect.com/science/article/pii/0004370272900495>.
- Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (Nov. 2011). “Natural Language Processing (Almost) from Scratch”. In: *Journal of Machine Learning Research* 12.null, 2493–2537. ISSN: 1532-4435.
- Conati, Cristina and Heather Maclaren (2009). “Empirically building and evaluating a probabilistic model of user affect”. In: *User Modeling and User-Adapted Interaction* 19.3, pp. 267–303.
- Cowen, Alan S and Dacher Keltner (2017). “Self-report captures 27 distinct categories of emotion bridged by continuous gradients”. In: *Proceedings of the national academy of sciences* 114.38, E7900–E7909.
- Cybenko, George (1989). “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of control, signals and systems* 2.4, pp. 303–314.
- Dahlbäck, Nils, Arne Jönsson, and Lars Ahrenberg (1993). “Wizard of Oz studies: why and how”. In: *Proceedings of the 1st International Conference on Intelligent User Interfaces*. IUI ’93. Orlando, Florida, USA: Association for Computing Machinery, 193–200. ISBN: 0897915569. DOI: 10.1145/169891.169968. URL: <https://doi.org/10.1145/169891.169968>.
- Dai, Andrew M and Quoc V Le (2015). “Semi-supervised Sequence Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf).
- Damasio, A. R. (1994). *Descartes’ error: Emotion, reason, and the human brain*. New York: Avon.
- Dan-Glauser, Elise S and Klaus R Scherer (2011). “The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance”. In: *Behavior research methods* 43.2, pp. 468–477.
- Darwin, Charles (1872). *The expression of the emotions in man and animals*. Vol. 3. John Murray London.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423/>.

- Dijk, Eric, Gerben van Kleef, Wolfgang Steinel, and Ilja Van Beest (Apr. 2008). "A Social Functional Approach to Emotions in Bargaining: When Communicating Anger Pays and When It Backfires". In: *Journal of personality and social psychology* 94, pp. 600–14. DOI: 10.1037/0022-3514.94.4.600.
- Dinucu-Jianu, David, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan (Nov. 2025). "From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng. Suzhou, China: Association for Computational Linguistics, pp. 272–292. ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.15. URL: <https://aclanthology.org/2025.emnlp-main.15/>.
- D'mello, Sidney and Art Graesser (2013). "AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2.4, pp. 1–39.
- Ekman, Paul (1971). "Universals and cultural differences in facial expressions of emotion". In: *Nebraska Symposium on Motivation* 19, pp. 207–283. ISSN: 0146-7875(Print).
- Ekman, Paul (1992b). "An argument for basic emotions". In: *Cognition and Emotion*, pp. 169–200.
- Ekman, Paul (1992a). "An argument for basic emotions". In: *Cognition & emotion* 6.3-4, pp. 169–200.
- Fadaee, Marzieh, Arianna Bisazza, and Christof Monz (July 2017). "Data Augmentation for Low-Resource Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 567–573. DOI: 10.18653/v1/P17-2090. URL: <https://aclanthology.org/P17-2090/>.
- Felbo, Bjarke, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann (Sept. 2017). "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1615–1625. DOI: 10.18653/v1/D17-1169. URL: <https://aclanthology.org/D17-1169/>.
- Feng, Shutong, Hsien chin Lin, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić (2025). *Emotionally Intelligent Task-oriented Dialogue Systems: Architecture, Representation, and Optimisation*. arXiv: 2507.01594 [cs.CL]. URL: <https://arxiv.org/abs/2507.01594>.
- Feng, Shutong, Hsien-chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Matthias Ruppik, Renato Vukovic, and Milica Gašić (Sept. 2024a). "Infusing Emotions into Task-oriented Dialogue Systems: Understanding, Management, and Generation". In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani. Kyoto, Japan: Association for Computational Linguistics, pp. 699–717. DOI: 10.18653/v1/2024.sigdial-1.60. URL: <https://aclanthology.org/2024.sigdial-1.60/>.
- Feng, Shutong, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić (June 2022). "EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4096–4113. URL: <https://aclanthology.org/2022.lrec-1.436/>.
- Feng, Shutong, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gašić (Sept. 2023). "From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue". In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani. Prague, Czechia: Association for Computational Linguistics, pp. 85–103. DOI: 10.18653/v1/2023.sigdial-1.8. URL: <https://aclanthology.org/2023.sigdial-1.8/>.

- Feng, Shutong, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gašić (Sept. 2024b). "Affect Recognition in Conversations Using Large Language Models". In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani. Kyoto, Japan: Association for Computational Linguistics, pp. 259–273. DOI: 10.18653/v1/2024.sigdial-1.23. URL: <https://aclanthology.org/2024.sigdial-1.23/>.
- Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, p. 378.
- Forgas, Joseph (Jan. 1995). "Mood and judgment: The Affect Infusion Model (AIM)". In: *Psychological bulletin* 117, pp. 39–66. DOI: 10.1037/0033-2909.117.1.39.
- French, Geoff, Michal Mackiewicz, and Mark Fisher (2018). "Self-ensembling for visual domain adaptation". In: *International Conference on Learning Representations*. Vancouver, Canada. URL: <https://openreview.net/forum?id=rkpoTaxA->.
- Frick, Robert (May 1985). "Communicating Emotion. The Role of Prosodic Features". In: *Psychological Bulletin* 97, pp. 412–429. DOI: 10.1037/0033-2909.97.3.412.
- Frijda, Nico H (1986). *The emotions*. Cambridge University Press.
- Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed (Sept. 2024). "Bias and Fairness in Large Language Models: A Survey". In: *Computational Linguistics* 50.3, pp. 1097–1179. DOI: 10.1162/coli\_a\_00524. URL: <https://aclanthology.org/2024.cl-3.8/>.
- Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky (2016). "Domain-adversarial training of neural networks". In: *Journal of machine learning research* 17.59, pp. 1–35.
- Gardner, Howard (1983). *Frames of mind: The theory of multiple intelligences*. Basic books.
- Gašić, M., C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young (2013). "On-line policy optimisation of Bayesian spoken dialogue systems via human interaction". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada, pp. 8367–8371. DOI: 10.1109/ICASSP.2013.6639297.
- Glas, Dylan F., Takashi Minato, Carlos T. Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro (2016). "ERICA: The ERATO Intelligent Conversational Android". In: *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 22–29. DOI: 10.1109/ROMAN.2016.7745086.
- Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. Ft. Lauderdale, FL, USA, pp. 315–323.
- Godfrey, John J. and Edward Holliman (1993). *Switchboard-1 Release 2*. LDC97S62. Philadelphia. URL: <https://www ldc.upenn.edu/>.
- Goodfellow, Ian J et al. (2013). "Challenges in representation learning: A report on three machine learning contests". In: *International conference on neural information processing*. Springer. Daegu, Korea, pp. 117–124.
- Gratch, Jonathan and Stacy Marsella (2004). "A domain-independent framework for modeling emotion". In: *Cognitive Systems Research* 5.4, pp. 269–306.
- Gratch, Jonathan et al. (May 2014). "The Distress Analysis Interview Corpus of human and computer interviews". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3123–3128. URL: [http://www.lrec-conf.org/proceedings/lrec2014/pdf/508\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf).
- Grice, HP (1975). "Logic and Conversation". In: *Syntax and Semantics: Vol. 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York: Academic Press, pp. 41–58. URL: <http://www.ucl.ac.uk/lis/studypacks/Grice-Logic.pdf>.
- He, Hua and Jimmy Lin (2016). "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement". In: *Proceedings of the 2016 conference of the north American chapter*

- of the Association for Computational Linguistics: human language technologies. San Diego, California, pp. 937–948.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. USA: IEEE Computer Society, 1026–1034. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.123. URL: <https://doi.org/10.1109/ICCV.2015.123>.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, Nevada, USA, pp. 770–778.
- He, Wanwei et al. (2022). “GALAXY: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-Supervised Learning and Explicit Policy Injection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Heck, Michael, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauer, Hsien-Chin Lin, and Milica Gašić (Nov. 2022). “Robust Dialogue State Tracking with Weak Supervision and Sparse Data”. In: *Transactions of the Association for Computational Linguistics* 10, pp. 1175–1192. ISSN: 2307-387X. DOI: 10.1162/tac1\_a\_00513. eprint: [https://direct.mit.edu/tac1/article-pdf/doi/10.1162/tac1\\_a\\_00513/2057218/tac1\\_a\\_00513.pdf](https://direct.mit.edu/tac1/article-pdf/doi/10.1162/tac1_a_00513/2057218/tac1_a_00513.pdf). URL: [https://doi.org/10.1162/tac1\\_a\\_00513](https://doi.org/10.1162/tac1_a_00513).
- Heck, Michael, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauer, Hsien-chin Lin, Carel van Niekerk, and Milica Gašić (July 2023). “ChatGPT for Zero-shot Dialogue State Tracking: A Solution or an Opportunity?” In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 936–950. DOI: 10.18653/v1/2023.acl-short.81. URL: <https://aclanthology.org/2023.acl-short.81/>.
- Hendrycks, Dan and Kevin Gimpel (2016). “Gaussian error linear units (gelus)”. In: *arXiv preprint arXiv:1606.08415*.
- Higashinaka, Ryuichiro, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro (2010). “Issues in Predicting User Satisfaction Transitions in Dialogues: Individual Differences, Evaluation Criteria, and Prediction Models”. In: *Spoken Dialogue Systems for Ambient Environments*. Ed. by Gary Geunbae Lee, Joseph Mariani, Wolfgang Minker, and Satoshi Nakamura. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 48–60. ISBN: 978-3-642-16202-2.
- Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.
- Høffding, Harald (1892). *Outlines of psychology*. Macmillan and Company, Limited.
- Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5, pp. 359–366.
- Hosseini-Asl, Ehsan, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher (2020). “A simple language model for task-oriented dialogue”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.
- Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). “Parameter-Efficient Transfer Learning for NLP”. In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR. Long Beach, CA, USA, pp. 2790–2799.
- Hu, Edward J, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2022). “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. Online. URL: <https://openreview.net/forum?id=nZeVKeFYf9>.

- Huang, Lei et al. (Jan. 2025). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Trans. Inf. Syst.* 43.2. ISSN: 1046-8188. DOI: 10.1145/3703155. URL: <https://doi.org/10.1145/3703155>.
- Ioffe, Sergey and Christian Szegedy (July 2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456. URL: <https://proceedings.mlr.press/v37/loff15.html>.
- Izard, Carroll E. (1977). *Human Emotions*. New York: Plenum Press.
- James, William (1884). "What is an Emotion?" In: *Mind* 9.34, pp. 188–205. ISSN: 00264423, 14602113. URL: <http://www.jstor.org/stable/2246769> (visited on 08/25/2025).
- Jiang, Jing and ChengXiang Zhai (June 2007). "Instance Weighting for Domain Adaptation in NLP". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Annie Zaenen and Antal van den Bosch. Prague, Czech Republic: Association for Computational Linguistics, pp. 264–271. URL: <https://aclanthology.org/P07-1034/>.
- Jurafsky, Daniel and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. URL: <https://web.stanford.edu/~jurafsky/slp3/>.
- Kaplan, Jared et al. (2020). "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361*.
- Khodak, Mikhail, Nikunj Saunshi, and Kiran Vodrahalli (May 2018). "A Large Self-Annotated Corpus for Sarcasm". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA). URL: <https://aclanthology.org/L18-1102/>.
- Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG]. URL: <https://arxiv.org/abs/1412.6980>.
- Konda, Vijay and John Tsitsiklis (1999). "Actor-Critic Algorithms". In: *Advances in Neural Information Processing Systems*. Ed. by S. Solla, T. Leen, and K. Müller. Vol. 12. Denver, CO, USA: MIT Press. URL: [https://proceedings.neurips.cc/paper\\_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf).
- Krippendorff, Klaus (1980). *Content Analysis: An Introduction to Its Methodology*. Sage publications.
- Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.
- Lange, Carl Georg (1885). "The mechanism of the emotions". In: *The classical psychologists*, pp. 672–684.
- Lazarus, Richard S (1991). *Emotion and adaptation*. Oxford University Press.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539. URL: <https://hal.science/hal-04206682>.
- Lee, Kyusong, Tiancheng Zhao, Alan W. Black, and Maxine Eskenazi (July 2018). "DialCrowd: A toolkit for easy dialog system assessment". In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Ed. by Kazunori Komatani, Diane Litman, Kai Yu, Alex Papangelis, Lawrence Cavedon, and Mikio Nakano. Melbourne, Australia: Association for Computational Linguistics, pp. 245–248. DOI: 10.18653/v1/W18-5028. URL: <https://aclanthology.org/W18-5028/>.
- Levin, Esther, Roberto Pieraccini, and Wieland Eckert (2002). "A stochastic model of human-machine interaction for learning dialog strategies". In: *IEEE Transactions on speech and audio processing* 8.1, pp. 11–23.
- Li, Xiang Lisa and Percy Liang (Aug. 2021). "Prefix-Tuning: Optimizing Continuous Prompts for Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353. URL: <https://aclanthology.org/2021.acl-long.353/>.

- Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu (Nov. 2017). "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Greg Kondrak and Taro Watanabe. Taipei, Taiwan: Asian Federation of Natural Language Processing, pp. 986–995. URL: <https://aclanthology.org/I17-1099/>.
- Lin, Hsien-Chin, Shutong Feng, Christian Geishauer, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasić (2023). "EmoUS: Simulating User Emotions in Task-Oriented Dialogues". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23*. Taipei, Taiwan: Association for Computing Machinery, 2526–2531. ISBN: 9781450394086. DOI: 10.1145/3539618.3592092. URL: <https://doi.org/10.1145/3539618.3592092>.
- Lin, Hsien-chin, Christian Geishauer, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gašić (Sept. 2022). "GenTUS: Simulating User Behaviour and Language in Task-oriented Dialogues with Generative Transformers". In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek. Edinburgh, UK: Association for Computational Linguistics, pp. 270–282. DOI: 10.18653/v1/2022.sigdial-1.28. URL: <https://aclanthology.org/2022.sigdial-1.28/>.
- Lin, Hsien-chin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauer, Michael Heck, Shutong Feng, and Milica Gasic (July 2021). "Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems". In: *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Haizhou Li et al. Singapore and Online: Association for Computational Linguistics, pp. 445–456. DOI: 10.18653/v1/2021.sigdial-1.47. URL: <https://aclanthology.org/2021.sigdial-1.47/>.
- Lin, Hsien-Chin, Benjamin Matthias Ruppik, Carel van Niekerk, Chia-Hao Shen, Michael Heck, Nurul Lubis, Renato Vukovic, Shutong Feng, and Milica Gašić (2025). *Prompt reinforcing for long-term planning of large language models*. arXiv: 2510.05921 [cs.CL]. URL: <https://arxiv.org/abs/2510.05921>.
- Liu, Siyang, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang (Aug. 2021). "Towards Emotional Support Dialog Systems". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 3469–3483. DOI: 10.18653/v1/2021.acl-long.269. URL: <https://aclanthology.org/2021.acl-long.269/>.
- Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. New Orleans, LA, USA. URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Luneski, Andrej, Panagiotis D Bamidis, and Madga Hitoglou-Antoniadou (2008). "Affective computing and medical informatics: state of the art in emotion-aware medical applications." In: *Studies in health technology and informatics* 136, p. 517.
- Maas, Andrew L, Awni Y Hannun, Andrew Y Ng, et al. (2013). "Rectifier nonlinearities improve neural network acoustic models". In: *International conference on machine learning*. Vol. 30. 1. PMLR. Atlanta, USA, p. 3.
- Manning, Christopher and Hinrich Schutze (1999). *Foundations of statistical natural language processing*. MIT press.
- Marcinkiewicz, Mary Ann (1994). "Building a large annotated corpus of English: The Penn Treebank". In: *Using Large Corpora* 273, p. 31.
- Marsden, Jerrold E and Anthony Tromba (2003). *Vector calculus*. Macmillan.
- Marsella, Stacy, Jonathan Gratch, Paolo Petta, et al. (2010). "Computational models of emotion". In: *A Blueprint for Affective Computing-A sourcebook and manual* 11.1, pp. 21–46.

- Mayer, John and Peter Salovey (1997). "What is Emotional Intelligence?" In: *P. Salovey & D. Sluyter (Eds). Emotional development and emotional intelligence Implications for educators*, pp. 3–31.
- Mehrabian, Albert (1996). "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". In: *Current psychology* 14.4, pp. 261–292.
- Mehrabian, Albert and James A Russell (1974). *An approach to environmental psychology*. the MIT Press.
- Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *International Conference on Learning Representations*. Scottsdale, Arizona, USA. URL: <https://api.semanticscholar.org/CorpusID:5959482>.
- Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko (June 2018). "SemEval-2018 Task 1: Affect in Tweets". In: *Proceedings of the 12th International Workshop on Semantic Evaluation*. Ed. by Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 1–17. DOI: 10.18653/v1/S18-1001. URL: <https://aclanthology.org/S18-1001/>.
- Mohammad, Saif and Peter Turney (June 2010). "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon". In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Ed. by Diana Inkpen and Carlo Strapparava. Los Angeles, CA: Association for Computational Linguistics, pp. 26–34. URL: <https://aclanthology.org/W10-0204/>.
- Mohapatra, Biswesh, Seemab Hassan, Laurent Romary, and Justine Cassell (May 2024). "Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 3967–3977. URL: <https://aclanthology.org/2024.lrec-main.352/>.
- Nekvinda, Tomáš and Ondřej Dušek (Aug. 2021). "Shades of BLEU, Flavours of Success: The Case of MultiWOZ". In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, pp. 34–46. DOI: 10.18653/v1/2021.gem-1.4. URL: <https://aclanthology.org/2021.gem-1.4>.
- Nickolls, John, Ian Buck, Michael Garland, and Kevin Skadron (2008). "Scalable parallel programming with CUDA". In: *ACM SIGGRAPH 2008 Classes*. SIGGRAPH '08. Los Angeles, California: Association for Computing Machinery. ISBN: 9781450378451. DOI: 10.1145/1401132.1401152. URL: <https://doi.org/10.1145/1401132.1401152>.
- Niekerk, Carel van, Andrey Malinin, Christian Geishausser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gašić (Nov. 2021). "Uncertainty Measures in Neural Belief Tracking and the Effects on Dialogue Policy Performance". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7901–7914. DOI: 10.18653/v1/2021.emnlp-main.623. URL: <https://aclanthology.org/2021.emnlp-main.623/>.
- OpenAI (2022). *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>.
- Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge University Press. DOI: 10.1017/CB09780511571299.
- Ouyang, Long et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Orleans, Louisiana, USA: Curran Associates, Inc., pp. 27730–27744. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf).

- Paek, Tim and Roberto Pieraccini (2008). "Automating spoken dialogue management design using machine learning: An industry perspective". In: *Speech Communication* 50.8, pp. 716–729. ISSN: 0167-6393. DOI: <https://doi.org/10.1016/j.specom.2008.03.010>. URL: <https://www.sciencedirect.com/science/article/pii/S0167639308000344>.
- Paiva, Ana, Joao Dias, Daniel Sobral, Ruth Aylett, Sarah Woods, Lynne Hall, and Carsten Zoll (2005). "Learning by feeling: Evoking empathy with synthetic characters". In: *Applied Artificial Intelligence* 19.3-4, pp. 235–266.
- Pan, Sinno Jialin and Qiang Yang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.
- Pang, Bo and Lillian Lee (June 2005). "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ed. by Kevin Knight, Hwee Tou Ng, and Kemal Oflazer. Ann Arbor, Michigan, USA: Association for Computational Linguistics, pp. 115–124. DOI: 10.3115/1219840.1219855. URL: <https://aclanthology.org/P05-1015/>.
- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (July 2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Philadelphia, PA, USA: Association for Computational Linguistics, pp. 79–86. DOI: 10.3115/1118693.1118704. URL: <https://aclanthology.org/W02-1011/>.
- Parkinson, Brian (1996). "Emotions are social." In: *British journal of psychology (London, England : 1953)* 87 ( Pt 4).4, pp. 663–683.
- Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (June 2013). "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, pp. 1310–1318. URL: <https://proceedings.mlr.press/v28/pascanu13.html>.
- Peng, Baolin, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao (Nov. 2020). "Few-shot Natural Language Generation for Task-Oriented Dialog". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 172–182. DOI: 10.18653/v1/2020.findings-emnlp.17. URL: <https://aclanthology.org/2020.findings-emnlp.17/>.
- Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: <https://aclanthology.org/D14-1162/>.
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: <https://aclanthology.org/N18-1202/>.
- Picard, Rosalind W. (1997). *Affective Computing*. Cambridge, MA: MIT Press. ISBN: 978-0-262-16170-1.
- Pichora-Fuller, M. Kathleen and Kate Dupuis (2020). *Toronto emotional speech set (TESS)*. Version V1. DOI: 10.5683/SP2/E8H2MF. URL: <https://doi.org/10.5683/SP2/E8H2MF>.
- Plutchik, Robert (1980). "A general psychoevolutionary theory of emotion". In: *Theories of emotion 1*, pp. 3–31.
- Prendinger, Helmut and Mitsuru Ishizuka (2005). "THE EMPATHIC COMPANION: A CHARACTER-BASED INTERFACE THAT ADDRESSES USERS' AFFECTIVE STATES". In: *Applied artificial intelligence* 19.3-4, pp. 267–285.

- Qi, Xianyu, Wei Wang, Lei Guo, Mingbo Li, Xiaoyu Zhang, and Ran Wei (2019). "Building a plutchik's wheel inspired affective model for social robots". In: *Journal of Bionic Engineering* 16.2, pp. 209–221.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018). "Improving language understanding by generative pre-training". In.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140, pp. 1–67.
- Raux, Antoine, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskenazi (2005). "Let's go public! taking a spoken dialog system to the real world". In: *Interspeech 2005*. Lisbon, Portugal, pp. 885–888. DOI: 10.21437/Interspeech.2005-399.
- Robert, Donovan and Rossiter John (1982). "Store atmosphere: an environmental psychology approach". In: *Journal of retailing* 58.1, pp. 34–57.
- Rosenthal, Sara, Noura Farra, and Preslav Nakov (Aug. 2017). "SemEval-2017 Task 4: Sentiment Analysis in Twitter". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Ed. by Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens. Vancouver, Canada: Association for Computational Linguistics, pp. 502–518. DOI: 10.18653/v1/S17-2088. URL: <https://aclanthology.org/S17-2088/>.
- Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986a). "Learning internal representations by error propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 318–362. ISBN: 026268053X.
- Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986b). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.
- Russell, J.A. (1980). "A circumplex model of affect". In: *Journal of personality and social psychology* 39.6, pp. 1161–1178. ISSN: 0022-3514.
- Russell, James A (1994). "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies." In: *Psychological bulletin* 115.1, p. 102.
- Saeki, Mao, Hiroaki Takatsu, Fuma Kurata, Shungo Suzuki, Masaki Eguchi, Ryuki Matsuura, Kotaro Takizawa, Sadahiro Yoshikawa, and Yoichi Matsuyama (Sept. 2024). "IntelLA: Intelligent Language Learning Assistant for Assessing Language Proficiency through Interviews and Roleplays". In: *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani. Kyoto, Japan: Association for Computational Linguistics, pp. 385–399. DOI: 10.18653/v1/2024.sigdial-1.34. URL: <https://aclanthology.org/2024.sigdial-1.34/>.
- Salovey, Peter and John D. Mayer (1990). "Emotional Intelligence". In: *Imagination, Cognition and Personality* 9.3, pp. 185–211. DOI: 10.2190/DUGG-P24E-52WK-6CDG. eprint: <https://doi.org/10.2190/DUGG-P24E-52WK-6CDG>. URL: <https://doi.org/10.2190/DUGG-P24E-52WK-6CDG>.
- Scherer, Klaus R (2005). "What are emotions? And how can they be measured?" In: *Social science information* 44.4, pp. 695–729.
- Schlossberg, Harold (1954). "Three dimensions of emotion." In: *Psychological review* 61.2, p. 81.
- Schmidt, Philip, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven (Sept. 2019). "Wearable-Based Affect Recognition-A Review". en. In: *Sensors (Basel)* 19.19.
- Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov (2017). "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347*.
- Searle, John R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 86–96. DOI: 10.18653/v1/P16-1009. URL: <https://aclanthology.org/P16-1009/>.
- Skantze, Gabriel (2007). *Error handling in spoken dialogue systems-managing uncertainty, grounding and miscommunication*. Gabriel Skantze.

- Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts (2013). "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. Seattle, Washington, USA, pp. 1631–1642.
- Sousa, Samuel and Roman Kern (2023). "How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing". In: *Artificial Intelligence Review* 56.2, pp. 1427–1492.
- Stevanovic, Melisa and Anssi Peräkylä (Apr. 2015). "Experience sharing, emotional reciprocity, and turn-taking". In: *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.00450.
- Strapparava, Carlo and Rada Mihalcea (2007). "Semeval-2007 task 14: Affective text". In: *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. Prague, Czech Republic, pp. 70–74.
- Stricker, Armand and Patrick Paroubek (2023). "Enhancing Task-Oriented Dialogues With Chitchat: A Comparative Study Based on Lexical Diversity And Divergence". In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. URL: <https://api.semanticscholar.org/CorpusID:265445499>.
- Sun, Guangzhi, Shutong Feng, Dongcheng Jiang, Chao Zhang, Milica Gasic, and Phil Woodland (Aug. 2024). "Speech-based Slot Filling using Large Language Models". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 6351–6362. DOI: 10.18653/v1/2024.findings-acl.379. URL: <https://aclanthology.org/2024.findings-acl.379/>.
- Sun, Guangzhi, Xiao Zhan, Shutong Feng, Philip C Woodland, and Jose Such (2025). "CASE-Bench: Context-Aware SafEty Benchmark for Large Language Models". In: *ICML*.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to sequence learning with neural networks". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 3104–3112.
- Sutton, Richard S, David McAllester, Satinder Singh, and Yishay Mansour (1999). "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Solla, T. Leen, and K. Müller. Vol. 12. Denver, CO, USA: MIT Press. URL: [https://proceedings.neurips.cc/paper\\_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf).
- Team, Gemini (2024). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv: 2312.11805 [cs.CL]. URL: <https://arxiv.org/abs/2312.11805>.
- Traum, David R (1999). "Computational models of grounding in collaborative systems". In: *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*. North Falmouth, Massachusetts, USA, pp. 124–131.
- Traum, David Rood (1995). *A computational theory of grounding in natural language conversation*. University of Rochester.
- Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460.
- Ultes, Stefan, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić, and Steve Young (2017). "Domain-Independent User Satisfaction Reward Estimation for Dialogue Policy Learning". In: *Interspeech 2017*, pp. 1721–1725. DOI: 10.21437/Interspeech.2017-1032.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Long Beach, California, USA: Curran Associates, Inc. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).
- Velásquez, Juan D. (1997). "Modeling emotions and other motivations in synthetic agents". In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on*

- Innovative Applications of Artificial Intelligence*. AAAI'97/IAAI'97. Providence, Rhode Island: AAAI Press, 10–15. ISBN: 0262510952.
- Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, and Alicia Abella (July 1997). "PARADISE: A Framework for Evaluating Spoken Dialogue Agents". In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, pp. 271–280. DOI: 10.3115/976909.979652. URL: <https://aclanthology.org/P97-1035/>.
- Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *The Eleventh International Conference on Learning Representations, ICLR 2023*. Kigali, Rwanda: OpenReview.net. URL: <https://openreview.net/forum?id=1PL1NIMMrw>.
- Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert (2013). "Norms of valence, arousal, and dominance for 13,915 English lemmas". In: *Behavior research methods* 45.4, pp. 1191–1207.
- Watkins, Christopher JCH and Peter Dayan (1992). "Q-learning". In: *Machine learning* 8.3, pp. 279–292.
- Watson, David and Auke Tellegen (1985). "Toward a consensual structure of mood." In: *Psychological bulletin* 98.2, p. 219.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022a). "Chain-of-thought prompting elicits reasoning in large language models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22*. Red Hook, NY, USA: Curran Associates Inc. ISBN: 9781713871088.
- Wei, Jason et al. (2022b). "Emergent Abilities of Large Language Models". In: *Transactions on Machine Learning Research*. Survey Certification. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=yzkSU5zdwD>.
- Weizenbaum, Joseph (Jan. 1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Commun. ACM* 9.1, 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. URL: <https://doi.org/10.1145/365153.365168>.
- Wen, Tsung-Hsien, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young (Apr. 2017). "A Network-based End-to-End Trainable Task-oriented Dialogue System". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 438–449. URL: <https://aclanthology.org/E17-1042/>.
- Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. URL: <http://aclweb.org/anthology/N18-1101>.
- Williams, Ronald J (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8.3, pp. 229–256.
- Winograd, Terry (1972). "Understanding natural language". In: *Cognitive psychology* 3.1, pp. 1–191.
- Xie, Qizhe, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le (2020a). "Unsupervised data augmentation for consistency training". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20*. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.
- Xie, Qizhe, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le (2020b). "Unsupervised data augmentation for consistency training". In: *Advances in neural information processing systems* 33, pp. 6256–6268.
- Xing, Bohao, Xin Liu, Guoying Zhao, Chengyu Liu, Xiaolan Fu, and Heikki Kälviäinen (2025). "EmotionHallucener: Evaluating Emotion Hallucinations in Multimodal Large Language Models". In: *ArXiv abs/2505.11405*. URL: <https://api.semanticscholar.org/CorpusID:278715371>.

- Yannakakis, Georgios N and John Hallam (2007). "Towards optimizing entertainment in computer games". In: *Applied Artificial Intelligence* 21.10, pp. 933–971.
- Yarowsky, David (June 1995). "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods". In: *33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, Massachusetts, USA: Association for Computational Linguistics, pp. 189–196. DOI: 10.3115/981658.981684. URL: <https://aclanthology.org/P95-1026/>.
- Young, Steve (2002). "Talking to machines (statistically speaking)". In: *7th International Conference on Spoken Language Processing (ICSLP 2002)*. Denver, Colorado, USA, pp. 9–16. DOI: 10.21437/ICSLP.2002-2.
- Young, Steve, Milica Gašić, Blaise Thomson, and Jason D. Williams (2013). "POMDP-Based Statistical Spoken Dialog Systems: A Review". In: *Proceedings of the IEEE* 101.5, pp. 1160–1179. DOI: 10.1109/JPROC.2012.2225812.
- Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). "Character-level convolutional networks for text classification". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 649–657.
- Zhou, Hao, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu (2018). "Emotional chatting machine: emotional conversation generation with internal and external memory". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press. ISBN: 978-1-57735-800-8.
- Zhu, Qi et al. (Dec. 2023). "ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Yansong Feng and Els Lefever. Singapore: Association for Computational Linguistics, pp. 106–123. DOI: 10.18653/v1/2023.emnlp-demo.9. URL: <https://aclanthology.org/2023.emnlp-demo.9/>.