



THE HONG KONG
POLYTECHNIC UNIVERSITY

香港理工大學

Pao Yue-kong Library

包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

EMOTIONALLY INTELLIGENT
CONVERSATIONAL AGENTS: FROM
UNDERSTANDING TO INTERACTION

JIASHUO WANG

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Computing

Emotionally Intelligent Conversational Agents: from
Understanding to Interaction

Jiashuo WANG

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

May 2025

Certificate of Originality

I hereby declare that this thesis is my work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

_____ (Signed)

 Jiashuo Wang (Name of student)

ABSTRACT

To advance artificial intelligence, it is essential to equip machines with emotional intelligence, thereby enhancing human-AI communication and relationships. In this thesis, I present our work on building emotionally intelligent conversational agents, focusing on three key themes: empathetic understanding, reliable responding, and engaging interaction.

For **empathetic understanding**, we propose two models, GREC and CARE, which are designed to generate empathetic responses by interpreting user emotions and the underlying emotional causalities through graph-based structures. While GREC reasons over an external commonsense knowledge graph, CARE integrates causal relationship inference directly within the model. To ensure **reliable responding**, we address two challenges. First, we introduce d-PM, a method to learn user preferences while accounting for individual disagreements, and align conversational agents accordingly. Second, to mitigate unhelpful responses that could hinder emotional support, we propose Muffin, a framework that reduces the likelihood of such responses by leveraging multi-faceted AI feedback. These two works are complementary, where one increases user satisfaction and the other mitigates unhelpfulness. Both methods are model-agnostic and can enhance transformer-based models, including state-of-the-art ones. The last theme centers on **engaging interaction** in emotionally intelligent conversational agents. We present two works: one for evaluation and one for model alignment. Since conversation engagement reflects the overall experience of an entire dialogue and involving real human users can be costly, we employ model-simulated users in our studies. First, we propose ClientCAST, a

framework to evaluate LLM-based therapists. After interacting with the conversational agents, simulated clients complete questionnaires to assess the overall conversational engagement. Second, to enhance engagement, we align conversational agents with conversations that are likely to produce higher engagement levels. This is achieved through Monte Carlo Tree Search for interaction, which identifies dialogue trajectories associated with greater user engagement.

Together, these contributions offer a comprehensive approach to building emotionally intelligent conversational agents that are empathetic, reliable, and engaging.

Publications Arising from the Thesis

Journal Papers

1. Jiashuo Wang, Wenjie Li, Peiqin Li, and Feiteng Mu. *GREC: Empathetic response generation through graph-based multi-hop reasoning on emotional causality*. Knowledge-Based Systems 233 (2021): 107547.

Conference Papers

2. Jiashuo Wang, Yi Cheng, and Wenjie Li. *CARE: Causality Reasoning for Empathetic Responses by Conditional Graph Generation*. Findings of the Association for Computational Linguistics: EMNLP 2022.
3. Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. *Aligning language models with human preferences via a bayesian approach*. Advances in Neural Information Processing Systems 36 (2023): 49113-49132.
4. Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. *Muffin: Mitigating Unhelpfulness in Emotional Support Conversations with Multifaceted AI Feedback*. Findings of the Association for Computational Linguistics: ACL 2024.
5. Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. *Towards a client-centered assessment of llm therapists by client simulation*. [Preprint]

Acknowledgements

I would like to express my deepest respect and gratitude to my supervisor, Professor Maggie Li, for her invaluable guidance, patience, and unwavering support throughout my Ph.D. journey. Her insightful advice and encouragement have been instrumental in shaping my research and academic growth. I am truly grateful for the opportunity to learn from her expertise and mentorship.

I would also like to extend my appreciation to my mates, friends, and senior graduates who have accompanied and helped me on this journey. Their support, discussions, and camaraderie have made this experience both intellectually stimulating and personally enriching. Special thanks to my family and my pattern for their unconditional love and encouragement, which have been a constant source of strength and motivation.

On a personal note, while the Ph.D. journey has been a significant chapter in my life—comprising 18.5% of my years so far—it does not define my entire future. I have cherished every moment of this experience, yet I remind myself that my future will be shaped by the choices I make and the efforts I continue to put forth. This journey has been a meaningful milestone, but it is only one step in the lifelong path of growth and discovery.

Finally, I am grateful for all the challenges and lessons that have helped me evolve, both as a researcher and as an individual. With this, I look forward to embracing the future with curiosity, determination, and an open heart.

Table of Contents

Abstract	ii
Publications Arising from the Thesis	iv
Acknowledgements	v
Table of Contents	viii
List of Tables	x
List of Figures	xii
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Thesis Outline	5
1.4 Contributions	12
2 Literature Review	14
2.1 Tasks and Solutions	14
2.1.1 Emotional Response	15
2.1.2 Empathetic Dialogue	17
2.1.3 Emotional Support	20
2.1.4 MultiModal Researches	23
2.2 Challenges	23
2.2.1 Dataset Collection	24
2.2.2 Evaluation	27
2.3 Applications	31
2.3.1 Emotional Support Chatbots	31
2.3.2 Human Trainer Agents	32
2.3.3 Human-AI Collaboration	32
I Empathetic Understanding	33
3 Understanding with Multi-hop over Knowledge Graph	35

3.1	Introduction	35
3.2	Method	36
3.2.1	Emotional Causality Reasoning and Encoding	37
3.2.2	Conversation Context Encoding	41
3.2.3	Emotional Causality Augmented Decoding	42
3.2.4	Training Objectives	45
3.3	Experiments	46
3.3.1	Experimental Settings	46
3.3.2	Evaluation	48
3.3.3	Further Analyses	51
3.4	Conclusion	56
4	Emotional Causalities Generation for Understanding	58
4.1	Introduction	58
4.2	Method	60
4.2.1	Graph Construction	61
4.2.2	Conditional Variational Graph Auto-Encoder	62
4.2.3	Graph-Infused Response Generation	64
4.2.4	Training Objectives	66
4.3	Experiments	66
4.3.1	Experimental Settings	66
4.3.2	Evaluation	68
4.3.3	Further Analyses	71
4.4	Conclusion	74
II	Reliable Responding	75
5	User Satisfactory Improvement via a Bayesian Approach	77
5.1	Introduction	77
5.2	Method	80
5.2.1	Problem Formulation	80
5.2.2	Preference Modeling with Disagreement	82
5.2.3	Calibration for Alignment	83
5.3	Experiments	85
5.3.1	Experimental Settings	85
5.3.2	Evaluation	90
5.3.3	Further Analyses	93
5.4	Conclusion	97
6	Unhelpfulness Mitigation via Multifaceted AI Feedbacks	99
6.1	Introduction	99
6.2	Method	101
6.2.1	Problem Formulation	101

6.2.2	Multifaceted AI Feedback	103
6.2.3	Unhelpfulness Mitigation	104
6.3	Experiments	106
6.3.1	Experimental Settings	106
6.3.2	Evaluation	109
6.3.3	Further Analyses	116
6.4	Conclusion	118
III	Engaging Interaction	119
7	A Client-Centric Automatic Evaluation of LLM Therapists	121
7.1	Introduction	121
7.2	Method	123
7.2.1	Problem Formulation	123
7.2.2	Client Simulation	124
7.2.3	Client-centered Questionnaire Completion	126
7.3	Experiments	127
7.3.1	Experimental Settings	127
7.3.2	Reliability of Evaluation Framework	129
7.3.3	Evaluation	137
7.3.4	Further Analyses	139
7.4	Conclusion	140
8	Conclusion and Future Research	145
	References	150

List of Tables

1.1	Introduction: Overview of Research Works in this Thesis.	6
1.2	Introduction: Datasets Used in this Thesis.	12
2.1	Literature Review: Overview of Tasks for Emotionally Intelligent Conversational Agents.	15
2.2	Literature Review: Overview of datasets for emotionally intelligent conversational agents.	25
3.1	GREC: A conversation from EmpatheticDialogue.	47
3.2	GREC: Automatic evaluation results.	49
3.3	GREC: Human rating results.	50
3.4	GREC: Human A/B test results.	51
3.5	GREC: Automatic evaluation results of ablation study.	52
3.6	GREC: Human rating results of ablation study.	52
3.7	GREC: Case study.	55
4.1	CARE: Automatic evaluation results.	68
4.2	CARE: Human rating results.	70
4.3	CARE: Human A/B test results.	71
4.4	CARE: Automatic evaluation results of the ablation study.	71
4.5	CARE: Human rating results of the ablation study.	71
4.6	CARE: Case study.	74
5.1	Alignment with d-PM: The MI codes in MI-Dataset.	87
5.2	Alignment with d-PM: Automatic evaluation results.	90
5.3	Alignment with d-PM: Human evaluation results on ESConv.	92
5.4	Alignment with d-PM: Case study.	97
6.1	Muffin: Automatic evaluation results.	111
6.2	Muffin: Multifaceted AI feedback results.	114
6.3	Muffin: Ablation study results.	116
6.4	Muffin: Case study.	117
7.1	Consistency of simulated clients' apparent trait in new counseling sessions given specific psychological profiles.	134
7.2	ClientCAST: Evaluation of simulated clients.	134
7.3	ClientCAST: Statistics of simulated clients' verbal styles.	135

7.4	ClientCAST: LIWC analysis of explanations for self-reported feelings.	137
7.5	ClientCAST: LIWC analysis of therapist responses.	139
7.6	ClientCAST: Statistics of LLM therapists' verbal style.	139

List of Figures

- 1.1 Introduction: Development timeline of emotionally intelligent conversational agents. 2

- 3.1 GREC: Overview architecture. 37
- 3.2 GREC: An example of emotional causality graph construction. . . 40
- 3.3 GREC: A case of the emotional causality graph. 56

- 4.1 CARE: Causality reasoning for empathetic understanding. 58
- 4.2 CARE: Overview architecture.. . . . 60
- 4.3 CARE: Model performs with the number of selected relationships k increasing. 73

- 5.1 Alignment with d-PM: Different levels of user satisfaction. 78
- 5.2 Alignment with d-PM: Preference modeling with disagreement and alignment accordingly. 80
- 5.3 Alignment with d-PM: The maximum and minimum preference scores of 10 candidates generated via diverse beam search. 84
- 5.4 Alignment with d-PM: Model performances with different candidate numbers K 93
- 5.5 Alignment with d-PM: Comparison between alignment with RL and Ours. 94

- 6.1 Muffin: Examples of unhelpful responses generated by various emotional support conversation models. 101
- 6.2 Muffin: Overview architecture. 103
- 6.3 Muffin: The prompt to obtain Multifaceted AI Feedback. 104
- 6.4 Muffin: Multifaceted AI feedback performances. 110
- 6.5 Muffin: An example presented to annotators for evaluation. . . . 111
- 6.6 Muffin: Human A/B test results. 113

- 7.1 ClientCAST: Overview Architecture. 124
- 7.2 ClientCAST: A piece of an interaction between Llama 3-70B, simulated client, and therapist. 143
- 7.3 ClientCAST: A simulated client \times LLM therapist interaction. . . . 143
- 7.4 ClientCAST: The proportion of inconsistent simulated clients who exhibit a higher level of apparent traits. 144

7.5	ClientCAST: Session outcome, therapeutic alliance and self-reported feelings scores of high- and low-quality sessions.	144
7.6	ClientCAST: LLM therapist assessments on session outcome, therapeutic alliance, and self-reported feelings.	144

Chapter 1

Introduction

1.1 Background

Emotional intelligence is a vital capability that empowers individuals to recognize, understand, manage, and influence their own emotions and those of others. This skill is essential for fostering effective communication, empathy, and strong interpersonal relationships [114, 45]. It is required in a wide range of downstream applications, such as medical dialogue [36], counseling [104], and social chatting [170]. Since the creation of ELIZA [142] in the 1960s, integrating emotional intelligence into conversational agents has been a primary focus in both research and industry, as shown in Fig. 1.1. This evolution is evident in various conversational agents, from PARRY [26] and ALICE [132] to more recent advancements like XiaoIce [170] and Pi AI [2].

Early conversational agents, such as PARRY and ALICE, were developed using rule-based or retrieval-based methods. These traditional approaches relied on predefined rules, patterns, or a fixed response corpus, limiting the agents' ability to understand the context or generate truly dynamic responses [61].

With the advent of deep learning and language models, recent conversational agents have been built on these models, enabling agents to generate more natural, context-aware, and flexible conversations. Unlike earlier rule-based or

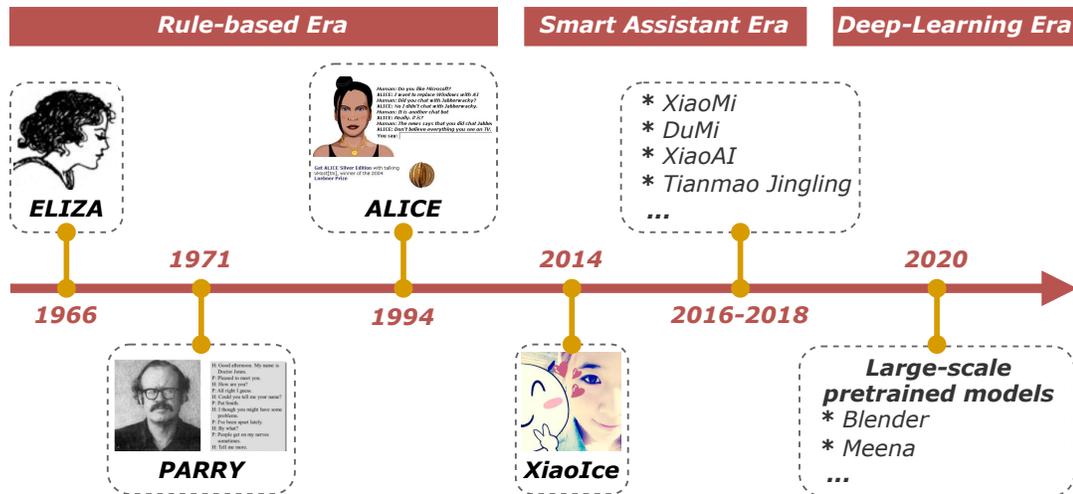


Figure 1.1: Introduction: Development timeline of emotionally intelligent conversational agents.

retrieval-based chatbots, modern AI-driven conversational agents can understand users' emotional and cognitive states, maintain context over longer interactions, and adapt their responses based on historical conversation and user states. These advances significantly improve the emotional intelligence of conversational agents. During the past six to seven years, numerous studies have focused on developing emotionally intelligent chatbots using these approaches. This thesis explores modern AI-driven methods to advance this field.

Emotionally intelligent conversational agents, implemented using neural networks, are at the forefront of enhancing human-computer interaction by recognizing, interpreting, and appropriately responding to human emotions. These advanced agents aim to foster empathetic, reliable, and engaging interaction, bridging the gap between human emotional complexity and artificial intelligence. When users share their feelings and thoughts, these AI models are designed to evaluate the emotional and cognitive states of the users. After the assessment process, conversational agents generate contextually appropriate responses that offer meaningful support and foster a sense of connection and

understanding, enhancing the overall user experience and strengthening the human-agent relationship. By continually refining the interaction, emotionally intelligent agents can provide mental health support to buffer daily stress [137] and even facilitate expert training [139].

Numerous studies have explored this subject. Some approaches analyze and incorporate user emotions into response generation [89, 82], others integrate external knowledge into the process [34, 113], and some use dialogue acts to guide the response generation [23]. Additionally, some methods combine at least two of the above-mentioned aspects for emotionally intelligent chatbots [164, 128]. While these approaches have made significant strides in enhancing the emotional intelligence of conversational agents, they often overlook the critical importance of a user-centric perspective.

1.2 Motivation

This thesis prioritizes the user's perspective in the design process of emotionally intelligent conversational agents. By focusing on this perspective, agents can gain a deeper understanding of the context and nuances of user emotions, leading to more meaningful empathy. Additionally, it enables agents to tailor their responses according to the user's unique preferences and behaviors, ultimately fostering a deeper connection and trust between the user and the agent. Moreover, a user-centric design enhances user satisfaction and engagement by aligning the agent's responses with the user's expectations. In essence, prioritizing user aspects is crucial for developing conversational agents that are not only intelligent but also supportive and empathetic in their interactions.

The goal of the work in this thesis is to develop innovative solutions for creating emotionally intelligent conversational agents by placing the user's per-

spective at the core of the design process.

The investigations are expected to provide emotional and mental health support with consideration of each user's specific situation, buffering their depression and stress in the real world. By doing so, these agents can ensure empathetic understanding, deliver reliable responses, and maintain engaging interactions. Therefore, this thesis is organized around three key themes step by step according to the emotional intelligence level of the agents: from empathetic understanding to engaging interaction. They are outlined as follows:

- **Empathetic Understanding:** How to effectively equip models with the capability to comprehend not only the user's emotions but also the experiences that trigger these emotions?
- **Reliable Responding:** How to enable models to generate reliable responses that are supportive from the user perspective while avoiding potentially harmful outputs?
- **Engaging Interaction:** How to develop models that facilitate genuinely engaging interaction, fostering deep connections with users, and providing effective support?

During my Ph.D. journey, I investigated these themes, generating innovative insights and proposing new methods. Notably, in the area of **Empathetic Understanding**, our work is among the first to incorporate the causes of emotions into conversation generation, enabling a more empathetic understanding of users. In terms of **Reliable Responding**, while many studies have focused on improving supportive response generation, our research uniquely emphasizes user feelings and prioritizes user perspectives. To the best of our knowledge, our work is also pioneering in evaluating and enhancing the effectiveness of emotionally intelligent conversational agents at the conversation level and

from a user perspective. These designs ultimately contribute to more **Engaging Interaction**.

1.3 Thesis Outline

Building on the three central themes discussed earlier, this thesis is structured into three parts. Each theme is explored in a dedicated part and illustrated through two research studies. An overview of these studies is provided in Table 1.1.

Part I explores the first research theme: **Empathetic Understanding**. Achieving empathetic understanding requires recognizing both the user’s emotional and cognitive states, which naturally equip agents with emotional intelligence from a user aspect. While emotional states correspond to the user’s emotions—previously studied for their role in enhancing empathetic response generation [89, 82]—cognitive states relate to the user’s experiences. However, not all user experiences contribute to empathetic understanding; those that are closely related to the user’s emotions are particularly important. For example, if a user says “I lost my wallet on my way to the theater,” focusing on “lost” would be more helpful to understand the user than the “theater.” Therefore, we would like to involve causes and effects of emotions to help understand the user. Specifically, we leveraged the graph structure to indicate the understanding process for their interpreter. Two works—W1 and W2—were presented in this thesis to show the effectiveness:

- In W1, we used the emotion causes reasoned from a causal knowledge graph to assist the conversation agents in understanding the user before response generation. Correspondingly, we designed an empathetic chatbot named GREC. Specifically, it is assumed that the latest user utterance

Research Theme	Work Index	Research Work	Publication Venue	Year
Empathetic Understanding	W1	Empathetic response generation through graph-based multi-hop reasoning on emotional causality	Knowledge-Based Systems	2021
	W2	CARE: Causality Reasoning for Empathetic Responses by Conditional Graph Generation	EMNLP	2022
Reliable Responding	W3	Aligning Language Models with Human Preferences via a Bayesian Approach	NeurIPS	2023
	W4	Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback	ACL	2024
Engaging Interaction	W5	Towards a client-centered assessment of LLM therapists by client simulation	Under Review	2024

Table 1.1: Introduction: Overview of Research Works in this Thesis.

suggests the user's emotions and feelings worth paying attention to, and the clauses in this utterance are regarded as emotion cues. Then, the emotional causes are identified in the previous historical conversation with a cause classifier. In order to figure out the causal relationship between the emotion cues and causes through a multi-hop reasoning over a causal knowledge graph, named ConceptNet [122], to infer semantic paths from keywords in the cause context to keywords in the emotion cues. These reasoned graphs are named emotional causality graphs. A multi-layer Graph Convolutional Network (GCN) is used to encode the graphs. The encoded graphs are used to serve for the response generation in two ways. On the one hand, the fused graph-aware representation is fed into the decoder to generate a generic vocabulary distribution for the empathetic response. On the other hand, to explicitly exploit the information of emotional causality, multi-hop reasoning on the emotional causality graphs is adopted to obtain the vocabulary distribution over the nodes of the graphs at each step of generation, which affects the final vocabulary distribution.

- W1 has demonstrated the effectiveness of incorporating emotional causes and effects for enhancing empathetic understanding and response generation. While it relies on an external reasoning process, our goal is to empower the model itself with the ability to internally reason about emotion-related causalities. Thus, W2 advanced the model's capabilities by enabling it to reason about the causal relationships among emotions and relevant events without relying on external knowledge graphs, achieving a deeper and more accurate empathetic understanding of users. Similar to W1, we also utilized a graph structure to show the emotional causalities given the context. To this end, we proposed a Conditional Varia-

tional Graph Auto-Encoder (CVGAE). This model utilizes the keywords mentioned in the conversation to represent the user’s experiences and an emotion word to denote the user’s emotional state. The CVGAE then reasons about the causal relationships among these elements, formulating a causal graph that captures all plausible causalities interdependently and simultaneously, conditioned on the user’s emotion, dialogue history, and predicted future dialogue content. This approach turns causality reasoning into a conditional graph generation task, enabling the model to infer not only explicitly stated causalities but also those that are plausible yet not directly expressed in the dialogue. Subsequently, a graph-infused decoder incorporates the information derived from this emotional causality graph into the response generation process through a multi-source attention mechanism. By integrating causality reasoning and response generation, we enhanced the model’s understanding of the user’s feelings and experiences. This integration allows the model to generate responses that reflect a more nuanced and empathetic understanding, thereby achieving truly empathetic responses without the need for external knowledge graphs.

In Part II, we explored the generation of **Reliable Responses**. While language models enhance the flexibility of text generation tasks, ensuring the reliability of their responses remains a critical challenge. Although numerous studies have proposed methods to improve the helpfulness of model-generated responses, especially in the emotional support context [164, 34, 23], few consider user-specific preferences for helpfulness. As a result, responses deemed “helpful” by the model may still fail to satisfy users. Moreover, enhancing helpfulness alone does not prevent the generation of harmful responses. A reliable conver-

sational system must also focus on mitigating unhelpfulness. To address these challenges in emotionally intelligent conversational systems, two studies (W3 & W4) explore methods to improve both the helpfulness and safety of generated responses.

- In W3, we focused on the language model alignment with users' preferences in terms of emotional support. While there can be inherent disagreements due to the subjective nature of human preferences, the language model is expected to try its best to satisfy each individual. Therefore, it is necessary for the language model to consider universal preferences through preference modeling. Previous approaches typically rely on majority voting or averaging to consolidate multiple inconsistent preferences into a merged one. However, such methods cannot capture the nuanced degrees of disaggregation among humans and may only represent a specialized subset of individuals, thereby lacking the ability to quantitatively disclose the universality of human preferences. Therefore, we proposed a novel Bayesian approach, which is designed to approximate a "universal preference" that comprises the preferences of "all individuals", given the preferences of several individuals. Specifically, the observed preference among selected individuals serves as prior knowledge. The Bayesian approach aims to leverage the distribution of all possible universal preferences (likelihood probability) to adjust and smooth the initially observed one, leading to the derivation of a universal preference (posterior). Upon obtaining the universal preference, the likelihood of the expected preference types is calculated to establish a preference score. This score is then utilized for further language model alignment to generate responses that can try the best to emotionally support each user.

- W4 was concentrated on unhelpfulness mitigation in the context of emotional support. Improving the response helpfulness does not guarantee reliable responses all the time, given the fact that approximately 30% of responses generated by the SOTA emotional support chatbot on the ES-Conv benchmark are identified as unhelpful under strict evaluation criteria. Their occurrence can significantly undermine supportive efforts and damage the trust between the users and the model. To address this critical issue, we introduced a novel model-agnostic framework called Muffin, which stands for Mitigating unhelpfulness with multifaceted AI feedback for emotional support. The Muffin framework tackles the problem through two key deliberations. First, it considers multiple facets of emotional support, recognizing that previous models often emphasize a single facet, such as empathetic expression or response coherence, potentially leading to unhelpfulness. To address this, Muffin incorporates a multifaceted AI feedback module that assesses responses from various emotional support facets. This module leverages the advanced capabilities of recent large language models, such as LLaMA, which are instruction-tuned to avoid inefficient and expensive human feedback collection. Second, Muffin directly minimizes the probability of generating unhelpful responses. While previous models typically optimize by minimizing the negative log-likelihood of golden responses, Muffin introduces an additional objective to specifically target and reduce the likelihood of unhelpful responses. This is achieved by contrasting unhelpful responses, identified by the feedback module, with non-unhelpful ones. Through these two steps, Muffin effectively mitigates unhelpful responses generated by emotional support conversation models. In addition, Muffin can be applied

on any transformer-based model, ensuring that a wide range of emotional support models can benefit from Muffin’s multifaceted feedback and unhelpfulness mitigation strategies.

In Part III, we would like to investigate the overall effect of the conversational agents from the interaction level. To the best of our knowledge, few previous works have paid attention to the interaction-level effect of emotionally intelligent conversational agents. This is because the user’s immediate reaction, which can significantly influence the agents’ performances during conversations, is difficult to obtain. Therefore, either evaluating or optimizing the emotionally intelligent conversational agents from a conversation level is a challenge. In part three, we take a small step towards the conversation-level evaluation. Accordingly, W5 was proposed:

- W5 was devoted to evaluating the overall performance of emotionally intelligent conversational agents during conversations from the users’ perspective. In particular, we utilized large language models (LLMs) to simulate clients in counseling therapy to assess the performance of AI therapists. This approach offers two significant advantages over involving real human participants: (1) Ethically, it mitigates the risk and unsafety that humans can frequently mimic clients and be exposed to potentially harmful AI-generated content. (2) Technically, it allows for consistent comparison of different AI therapists’ performances with the same client. To achieve this, we developed ClientCAST, a client-centered framework for automatically assessing AI therapists. In this framework, the simulated client interacts with an LLM therapist and completes questionnaires related to the interaction. The results of these questionnaires enable the assessment of AI therapists from three client-centered perspectives.

Research Theme	Work Index	Dataset
Empathetic Understanding	W1	EmpatheticDialog
	W2	
Reliable Responding	W3	Emotional Support Conversation
	W4	
Engaging Interaction	W5	High-Low Quality Counseling

Table 1.2: Introduction: Datasets Used in this Thesis.

Since the focus is different in different works, we adopt different datasets, as shown in Table 1.2. The detailed information is included in each part.

1.4 Contributions

The contributions of this thesis are summarized as follows:

- This thesis explores the design of emotionally intelligent conversational agents using deep learning techniques. we proposed user-centric approaches to enhance user understanding, response generation, and engagement in interactions.
- We were among the first to research the causes and effects of emotions in dialogue, with the goal of achieving a more empathetic understanding of the user’s state and crafting more empathetic responses. We proposed two distinct approaches that utilized graph-based representations to model the comprehension process effectively.
- This thesis introduced an important research direction focused on reliable responding. We developed methods to increase the supportiveness

of responses from the user's perspective while minimizing unhelpful responses that may harm users.

- To the best of our knowledge, the works in this thesis are the first to emphasize the overall interaction effect from a user perspective. We investigated methods to evaluate counseling conversations, which can be used to improve the engagement of conversational agents.

Chapter 2

Literature Review

This chapter reviews existing studies on emotionally intelligent conversational agents, with a primary focus on key tasks and the approaches used to address them. Additionally, dataset collection and evaluation remain significant challenges in developing emotionally intelligent chatbots, making them a relatively important topic for discussion. While emotionally intelligent conversational agents are primarily designed to provide emotional support to individuals in need, they can also assist supporters in their roles. Therefore, this chapter also explores their potential applications in the last section.

2.1 Tasks and Solutions

With research on emotionally intelligent chatbots, scholars and practitioners continue to deepen their understanding of emotional intelligence and its diverse applications. Additionally, advancements in underlying techniques have accelerated, shifting the focus of research. As a result, existing studies can be broadly categorized into three main areas: emotional response, empathetic dialogue, and emotional support. The following sections will outline the objectives of each task and discuss the various approaches that have been employed to address them. An overview of these tasks is shown in Table [2.1](#).

Task	Focus	Popular Time	Common Methods
Emotional Response	To generate responses that both accurately and naturally express a predefined emotion.	2017-2019	<ul style="list-style-type: none"> • emotional embeddings incorporation • emotion-rich decoding • emotional vocabulary computation
Empathetic Dialogue	To understand and predict users' emotional states, along with their experiences and thoughts.	2018-	<ul style="list-style-type: none"> • multi-task learning • external knowledge incorporation • emotional cause reasoning
Emotional Support	To guide users toward positive feelings and thoughts by buffering their depression and distress after empathetic understanding.	2021-	<ul style="list-style-type: none"> • user state reasoning • dialogue act planning • therapy technique implementation

Table 2.1: Literature Review: Overview of Tasks for Emotionally Intelligent Conversational Agents.

2.1.1 Emotional Response

Early research on emotionally intelligent chatbots primarily focused on emotional expression [166, 168, 121, 171]. This task, referred to as emotional response, involves producing model-generated responses that can naturally and accurately exhibit emotional tones. In addition, given a user query, most of these studies were aimed at the output that can respond to the user in semantics while accurately expressing a predefined emotion.

Limited to the model architecture, i.e., recursive neural network (RNN), existing works typically addressed this task in the following three approaches:

Emotional Embedding Incorporation Some methods incorporate emotional embeddings for emotional expression. Usually, an emotional embedding is concatenated with the semantic word embedding. Then, such information

plays an important role in the decoder hidden state computation or loss computation. The emotional embeddings can be obtained from existing emotional embeddings, such as the Valence Arousal & Dominance (VAD) [91] embedding, or be trained with the generation model.

Emotion-Rich Decoding Another straightforward way is to enrich the decoding process with emotional information. For instance, EmoDS designs extra attention to related emotional words [121]. In MojiTalk [171], authors adopt a reinforced CVAE to guide the response generation with the specific emotion as the condition. [43] investigates emotion embeddings and emoji expression for empathetic understanding and expression in booking dialogue systems. ECM uses an internal memory, which is connected with the gated recursive unit (GRU) using a write gate and a read gate, to indicate the emotion state decay [168].

Emotional Vocabulary Computation Leveraging emotional vocabulary to influence the final word selection is also a classical approach. Typically, the vocabulary distribution is computed with the final decoder hidden states. In order to emphasize emotional expression, another distribution of emotional vocabulary can also be involved to determine the final word probabilities.

In this task, the emotional expression of the model-generated responses is usually evaluated by the accuracy of the emotion of the response. In particular, an emotional classifier was adopted to classify the generated responses and then evaluate the effectiveness of the methods.

Simply expressing the emotion in the response is neither practical nor a true demonstration of emotional intelligence. Moreover, with advancements in model training techniques and the emergence of powerful language mod-

els, this task has become relatively straightforward and simple to address. As a result, recent research in the field of natural language processing has shown limited interest in this area.

2.1.2 Empathetic Dialogue

Empathy is the ability to perceive and understand an individual's feeling from their aspects. It involves not only the actual or inferred emotional states of the other person but also their experiences and thoughts [99, 32]. Accordingly, the task of empathetic dialogue aims to understand the user and then respond to acknowledge it appropriately. Unlike to the emotional response, empathetic dialogue emphasizes understanding of users.

A number of methods have been proposed to develop empathetic conversational agents, which can be divided into two categories according to their empathy type when understanding the users:

Affective Empathy This line of work is focused on the comprehension of the user's emotions inferred from the historical conversation before an appropriate response, which is referred to as affective empathy [120]. The simplest method is to predict the user's emotion(s) and then concatenate the emotion(s) and the user query as the model input for response generation [109]. However, it has been proven ineffective to utilize merely one emotion to understand the user and express empathy. It is because the user's emotional state can be complex, and one emotion is unable to present itself. MoEL proposes a method for blending emotional responses in dialogue generation [82]. Its architecture consists of a shared listener that captures the dialogue context and multiple emotional listeners, each generating responses with a specific emotion. The final response

is a weighted combination of the shared listener's output and the emotional listeners' outputs, where the weights are determined by the user's predicted emotion distribution based on the dialogue context. The model is trained using multitask learning, which involves maximizing the maximum likelihood estimator (MLE) for the ground truth response and the probability of the user's actual emotion. Based on MoEL, MIME was proposed to allow the response to mimic the user's emotion to show their understanding of the user [89]. The user's emotion is comprehended using polarity-based emotion (negative or positive) based on the dialogue context. Then, the emotion representations are considered in the response generation for the emotional mimicry. This approach works, since its design somehow achieves the effects of somatic empathy, which involves having a physical reaction in response to what someone else is experiencing [107]. EmpDG deepened the comprehension of user emotions with additional consideration of fine-grained token-level emotions [74]. In particular, when encoding the dialogue context, emotion-related words are encoded by an additional transformer encoder and concatenated after the dialogue context vectors. Furthermore, EmpDG is designed to enhance the understanding of the user's emotional state by involving emotional information in the next user utterances as feedback. Thus, the model is trained with an additional loss to enhance the model's empathy ability. Notably, all three aforementioned works used the embedding of the first input token, i.e., [CLS], to predict the user's actual emotion through multitask learning. Moreover, fine-tuning pretrained language models on an empathetic dialogue corpus with multiple training objectives is also a straightforward and effective method, as demonstrated by CAiRE [81].

Affective & Cognitive Empathy In addition to considering user emotions, understanding the user's experiences and thoughts enhances cognitive empathy, which involves grasping another person's mental state and predicting their potential reactions to a situation.

S1: I'm sorry. I understand that you're feeling sad and that what you're experiencing is difficult.

S2: I'm sorry to hear about your grandmother. I know you miss her, and I'm here for you.

For example, the statement **S1** demonstrates cognitive empathy, while **S2** reflects mainly affective empathy. A common and straightforward method to understand and predict the user's experiences and thoughts is to involve external knowledge to assist the understanding of the user's statements. Various external knowledge bases are involved, including semantic knowledge (WordNet), commonsense knowledge (ConceptNet, Wikidata, and YAGO), causal knowledge (Cause Effect Graph [79] and ATOMIC [56]), and sentimental knowledge (NRC-VAD [93] and SenticNet [19]). KEMP [75] adopts the NRC-VAD and ConceptNet to enrich the dialogue context by linking related concepts with words in it. Therefore, an emotional context graph is constructed. For emotional expression, KEMP learns emotional signal from this graph and uses the signal to guide the empathetic response generation. GREC [136] understands both the affective and cognitive states of the users by considering the causes of emotion. It first identifies the cause cues in the dialogue context that can trigger the user's emotion in the immediate turn. Then, emotional causality graphs are constructed based on the keywords that can represent the emotion and the causes by multi-hop reasoning over the Cause Effect Graph. These graphs are then used to enrich the model's empathetic expression. Furthermore,

some other works could enable the model itself to reason relevant knowledge to understand the user's cognitive state without extraction from external knowledge bases. For instance, CEM [113] uses COMET, a commonsense reasoning model trained in ATOMIC, to infer the user's feelings and thoughts (intention, needs, effect, wants). Then, both the dialogue context and the inferred knowledge are combined for the response generation. Some works focus on deeply exploring the user's experiences related to emotions within the dialogue context, without relying on external knowledge. For example, EMMA [77], RecEC [44], and GEE [63] identify the causes of emotions in the dialogue context and use this information to influence response generation. CARE [134] reasons both the causes and effects of user emotions using a causal graph generated by a conditional variational graph auto-encoder. Additional attention is then given to this causal graph during response decoding. Recently, large language models have demonstrated the ability to generate users' emotional and cognitive states through prompting [146] or chain-of-thought (CoT) fine-tuning.

2.1.3 Emotional Support

Emotional support aims to alleviate individuals' emotional suffering by helping them understand and confront their current challenges [16, 50]. It can take the form of assistance with daily distress and depression or counseling therapy. The task of emotional support conversations requires a deeper level of emotional intelligence than the aforementioned two tasks, with the added responsibility of managing users' emotions and guiding the conversation.

When developing emotional support conversational agents, the following factors are consistently considered in existing studies:

User States As mentioned in the previous section, incorporating external knowledge has been proven to effectively enhance empathetic understanding of users, which serves as a fundamental pillar of emotional support. By integrating relevant knowledge, models can better comprehend users' emotions, contextual nuances, and underlying concerns, benefiting from meaningful support. For example, KEMI [34] leverages COMET to generate relevant knowledge and constructs a graph with the support of HEAL, a mental health knowledge graph [145], to facilitate empathetic comprehension. Similarly, MISC [128] improves the model's comprehension of user posts by integrating commonsense knowledge inferred through COMET. Beyond incorporating external knowledge, the method proposed in [164] focuses on the transition between cognitive and emotional states among user utterances. Overall, these approaches align with techniques used in empathetic conversations. More recently, some methods have also prompted language models to reason about users' evolving emotional and cognitive states using instructions or chain-of-thought (CoT) reasoning, as demonstrated in [133] and [78].

Communication Skills While empathetic understanding is essential for emotional support, effectively guiding the conversation is equally important. To achieve this, various communication strategies are employed in the method design. MISC [128] introduces a mixed-strategy learning module to guide response generation, seamlessly integrating multiple strategies based on a distribution derived from the [CLS] token embeddings in the input dialogue context. Inspired by A* search, MultiESC [23] anticipates communication strategies by estimating future user feedback, selecting those that yield the most favorable long-term outcomes. KEMI [34]

prioritizes proactivity, dynamically switching between initiative-driven and non-initiative strategies to either provide empathetic responses or initiate problem-solving discussions as needed. TranESC [164] recognizes that responses can emphasize different aspects—semantic expression, emotional expression, or strategic intent—and allows for fluid transitions between these elements. In [135], the authors consider universal communication preferences and optimize the model to generate responses guided by strategies designed to accommodate diverse user needs. Muffin [137] also considers the effectiveness of communication skills by mitigating the likelihood of responses adopting inappropriate communication skills.

Therapy Techniques This line of research explores the use of artificial intelligence to implement therapy techniques for providing mental health support. Unlike emotional support for everyday issues such as depression, this type of support focuses on addressing cognitive distortions and facilitating behavior change. Among various therapeutic techniques, Cognitive Behavioral Therapy (CBT) is one of the most practical and effective treatments, helping individuals manage their problems by changing their thoughts and behaviors [11]. A common implementation method involves the use of large language model (LLM) agents. For instance, CBT-LLM [96] is trained on a psychological health QA dataset generated by GPT-3.5, which is prompted with instructions containing CBT knowledge. CoCoA [70] incorporates both a basic memory and a cognitive distortion memory to store personal information and identify cognitive distortions, respectively. For response generation, it employs dynamic prompting to determine the specific response act, effectively guiding the generation process. [156] proposed AutoCBT, an autonomous multi-agent pipeline, where a

counselor agent serves as the interface and several supervisor agents function as the backend. Specifically, a routing mechanism determines when the supervisors take on their roles, while memory mechanisms store historical context for these supervisors.

While previous studies have focused on improving response generation, the overall conversation-level support effect has received less attention. To the best of our knowledge, few methods—apart from our works, which will be discussed in Part III—have been proposed to address this issue.

2.1.4 MultiModal Researches

In addition to the text (utterance content), other information, such as images (facial information and stickers) and voices, contribute to emotional intelligence significantly. As a result, research on emotional intelligence conversational agents involving multiple modalities is always an important focus. For instance, there is work considering the emotional status contained in the voice and facial expression to implement a virtual coach [126]. [39] proposes an amazing open-source multi-modal empathetic chatbot, which adopts multi-modal encoders and decoders for understanding and generation, and an LLM for intermediate reasoning. In addition, images can assist emotional expression. For instance, PEGS is proposed to integrate stickers with textual interaction [162], which enables richer empathetic information.

2.2 Challenges

Building and evaluating emotionally intelligent conversational agents can be challenging for several reasons:

- Ethical and privacy concerns: Research in this domain often involves sensitive data, such as private conversations, mental health counseling, or personal experiences, raising significant ethical and privacy issues.
- Complexity of emotional intelligence: Emotional intelligence itself is a multifaceted and evolving concept, encompassing empathy, self-awareness, emotional regulation, and social skills. Defining and replicating these abilities in conversational agents remains a significant challenge.
- Lack of standardization: Emotional expression and perception vary across individuals, cultures, and contexts, making it difficult to establish universal benchmarks for emotional intelligence in conversational agents.

Beyond methodological challenges, dataset collection and evaluation methods also pose significant difficulties due to these factors. This section will discuss these two major challenges in the research of emotionally intelligent conversational agents.

2.2.1 Dataset Collection

Collecting dialogues that require significant emotional intelligence can be challenging. Previous studies have primarily relied on one of three methods for dataset collection: web crawling, crowdsourcing, or data synthesis. Table 2.2 provides an overview of existing datasets used in studies on emotionally intelligent conversational agents. The table includes the dataset name, source, scale, publication year, whether an official split is available, and whether the conversations are multi-turn interactions.

Web Public Data Social media platforms provide a valuable source of data for analyzing emotional expression. Prior studies, for example, have utilized

Table 2.2: Literature Review: Overview of datasets for emotionally intelligent conversational agents.

Dataset	Source	Scale		Year	Official Split	Multi Turn
		#Conv.	#Utter.			
Emotional STC (ESTC) [168]	Weibo	4.4m	8.8m	2018	Yes	No
Persona-based Empathetic Conversation (PEC) [159]	Reddit	355k	833k	2018	Yes	No
Empathetic Dialogues [109]	Crowdsourcing	25k	107k	2019	Yes	Yes
High-Low Quality Counseling [105]	YouTube	259	11k	2019	No	Yes
Emotional Support Conversation (ESC) [87]	Crowdsourcing	1k	31k	2019	Yes	Yes
efaqa [48]	-	20k	-	2020	No	Yes
XiaoAI Empathetic Conversation (X-EMAC) [77]	Online logs	17k	67k	2021	No	Yes
Emotional Dialogues in OpenSubtitles (EDOS) [147]	OpenSubtitles	1M	3M	2021	No	Yes
PsyQA [123]	Yixinli Q&A column	56k	112k	2021	No	No
expert-Annotated MI (AnnoMI) [151]	YouTube	133	-	2022	Non	Yes
SMILECHAT [108]	Synthetic	55k	2k	2023	No	Yes
SoulChatCorpus [22]	Synthetic	256k	2m	2023	No	Yes
CPsyCounR [158]	Synthetic	3.1k	24k	2024	No	Yes
PsyDTCorpus [154]	Synthetic	5k	91k	2024	Yes	Yes

dialogues (post-comment pairs) from Weibo [117], annotating the emotions expressed by speakers [168]. Similarly, Reddit serves as a rich data source, particularly for discussions on topics such as depression and mental health. The Persona-based Empathetic Conversation (PEC) dataset [167] was derived from two subreddits: *happy* and *offmychest*, with data filtered using specific rules. However, a key limitation of this method is that the dialogues from these platforms are single-turn interactions and lack real-time engagement. Additionally, responses are not based on familiarity with the original poster, as they stem from post-comment pairs rather than sustained conversations. Beyond social media platforms, data can be collected from other platforms. For instance, XiaoAI Empathetic Conversation (X-EMMA) [77] was collected from online logs and annotated with emotion classes and causes. PsyQA [124] was crawled from the Q&A column of Yixinli, a Chinese mental health service platform. YouTube also provides rich public counseling demonstrations. High-Low Quality Counseling dataset [105] and expert-Annotated MI (AnnoMI) dataset [151] were derived from public videos on YouTube. Emotional Dialogues in OpenSubtitles (EDOS) [147] was obtained from the OpenSubtitles [83], a public movie and TV subtitles corpus.

Crowdsourcing Many existing datasets are collected through crowdsourcing. For example, the Empathetic Dialogue dataset [109] and the Emotional Support Conversation (ESC) dataset [85] were created using Amazon Mechanical Turk (MTurk). Specifically, workers were hired and instructed to engage in conversations following predefined guidelines. In general, crowdsourced data is of high quality. However, due to the high costs associated with training and data collection, datasets obtained through this

method are typically small in scale.

Synthetic Data With the advancement of LLMs, many studies have leveraged LLMs to synthesize datasets. SMILE [108] used ChatGPT to rewrite public single-turn conversations into multi-turn dialogues, creating SMILECHAT. Similarly, SoulChatCorpus [22] employed a prompt with ChatGPT to transform single-turn psychological counseling conversations into multi-turn empathetic dialogues. Notably, the original single-turn conversations were obtained through crowdsourcing. In [158], the authors collected online psychological counseling reports and used them to construct CPsyCounR. Additionally, PsyDTCorpus [154] was generated using dynamic one-shot learning with GPT-4. Synthesizing datasets helps balance cost and quality, but human verification remains essential. Moreover, model-generated data may not always align with real-world distributions, posing challenges for practical applications.

Compared to datasets collected through crowdsourcing or synthesis, web-crawled datasets are often simpler in terms of emotional intelligence exploration and formatting but tend to be larger in scale.

2.2.2 Evaluation

When evaluating content generated by emotionally intelligent conversational agents, both semantic accuracy and emotional intelligence should be considered. Numerous methods have been proposed to assess semantic accuracy, with the most common approach being the comparison of generated responses with ground truth references. For instance, BLEU [100] and ROUGE [80] measure n-gram similarity, while BertScore [160] captures embedding similarity. Additionally, qualitative metrics such as Fluency and Consistency are often used to

evaluate responses. In particular, Fluency assesses the grammatical correctness and naturalness of a response, ensuring it reads smoothly and is free from linguistic errors. Consistency measures whether the response aligns logically with the given context and does not contradict prior statements or known facts. The diversity of responses is also taken into account by Dist-n [73].

However, evaluating emotional intelligence remains challenging. As mentioned earlier, emotional expression and perception vary across individuals and cultures, making standardized evaluation difficult. Additionally, emotional intelligence is inherently complex, further complicating the assessment process. Previous studies have proposed various evaluation methods, which will be discussed in this subsection.

Evaluation Methods The two most common methods for evaluating emotionally intelligent conversational agents are rating and comparison. In the rating approach, annotators assess an agent-generated response along with its corresponding context, scoring it based on a specific metric. In the comparison method, annotators evaluate two responses—typically generated by different models—for the same conversation context and determine which one performs better. Annotators can be classifiers, LLMs, or human evaluators.

Evaluation Metrics To assess emotional expression and perception effectively, various metrics are utilized:

- **Emotion Accuracy:** This metric evaluates whether the generated response correctly identifies and expresses the intended emotion in a given context. It measures alignment with expected emotional cues and appropriateness of emotional tone. This metric is mainly used in emotional response generation, such as in [168].

- **Strategy Accuracy:** This assesses whether the conversational agent employs the appropriate emotional strategies, such as comforting, encouraging, or sympathizing, based on the context and user's emotional state. This metric is adopted in works that involve support strategies to guide the response generation [23, 128]
- **Fact-Check / Knowledge:** It evaluates whether the response maintains factual correctness and does not introduce misinformation while addressing the emotional aspect of the conversation. Studies enhancing responses with external knowledge evaluate models using this metric [128].
- **Empathy:** This metric measures the agent's ability to understand and respond to the user's emotions in a way that demonstrates compassion and emotional awareness. It is widely used in various works [137, 34, 23, 128, 136].
- **Comforting:** This measures the agent's ability to soothe and reassure the user, particularly in distressing or emotionally sensitive situations. A comforting response should acknowledge the user's emotions and offer words of encouragement or support. This metric is used in works including [137, 23].
- **Suggestion:** It assesses whether the agent provides relevant and constructive advice, recommendations, or alternative perspectives that can help the user navigate their situation effectively. This metric is used in works including [137, 23].
- **Identification:** It evaluates the agent's ability to accurately recognize and reflect the user's emotional state, ensuring that the response is contextually appropriate and emotionally attuned. This metric is used in works including [137, 23].

- **Helpfulness:** This evaluates whether the response provides meaningful help. This is an overall of the effectiveness of support, guidance, or solutions that address the user’s concerns. This metric is used in works including [137].
- **Safety:** This metric evaluates whether the response is ethical, non-harmful, and avoids generating content that could be offensive, manipulative, or emotionally damaging [108].
- **Proactivity:** This measures the agent’s ability to anticipate user needs and offer assistance or emotional support before being explicitly asked, contributing to a more responsive interaction. This metric is used in works including [34].
- **Engaging:** This metric assesses whether the agent successfully engages the user in the conversation, ensuring that the user feels heard and encouraged to fully express their emotions by the end of the interaction.

The above metrics are commonly used in previous studies, while additional metrics with clear definitions can be proposed to evaluate conversational agents in different contexts. However, a standardized set of metrics that ensures consistency across evaluations—similar to how BLEU and ROUGE are widely adopted for semantic assessment—remains lacking for emotional intelligence evaluation.

Evaluation Pipelines Several evaluation pipelines have been proposed to assess the conversational behaviors of LLM therapists. For example, BOLT [24] is a proof-of-concept computational framework developed to systematically assess the conversational behaviors of LLM therapists. It seeks to compare their interactions with those of both high- and low-quality human therapists and explore

whether their responses can be adjusted to exhibit more desirable therapeutic behaviors. Similarly, ClientCAST [138] was introduced to assess the overall performance of LLM therapists in counseling conversations from the client's perspective. In ClientCAST, LLMs are prompted to assume the role of a client, engage in conversations with LLM therapists, and subsequently complete questionnaires evaluating their counseling experience.

2.3 Applications

Emotionally intelligent conversational agents are typically designed to provide emotional support to users. However, they can also assist emotional supporters by generating draft responses or helping train them. Therefore, the applications of these agents will be discussed in this subsection.

2.3.1 Emotional Support Chatbots

Most of the previous works, which have been mentioned in Section 2.1, focus on conversational agents that provide emotional support to users. These chatbots are designed to recognize and respond to users' emotions, offering empathetic interactions and psychological support. They can be integrated into mental health platforms, customer service, and social companionship applications. By leveraging natural language understanding and affective computing, these agents help users feel heard, validated, and comforted. Some advanced models use therapeutic techniques, such as CBT strategies, to guide users through emotional challenges.

2.3.2 Human Trainer Agents

In addition to directly supporting users, emotionally intelligent conversational agents can serve as training tools for human supporters, such as therapists and counselors. These agents can simulate emotionally charged conversations, allowing trainees to practice their responses in a controlled environment. By providing real-time feedback and analysis, these systems help improve emotional intelligence, active listening skills, and empathy in human trainees. Moreover, they can generate sample responses, assisting human supporters in refining their communication strategies. For instance, Patient- Ψ [139] has been proposed to provide simulated counseling scenarios for trainees and offer feedback on their performance after conversations.

2.3.3 Human-AI Collaboration

Emotionally intelligent conversational agents can also enhance human-AI collaboration by assisting human professionals in managing emotionally sensitive conversations. For example, in customer service scenarios, these agents can analyze customer sentiment and suggest appropriate responses to human agents in real-time. In mental health settings, AI-driven assistants can help therapists track patient emotions over time, identify patterns, and provide supplementary support during sessions [119] or even between sessions. By working alongside humans rather than replacing them, these agents create a synergistic relationship, leveraging both human intuition and AI-driven analysis to improve the overall quality of emotional support.

Part I

Empathetic Understanding

Responding with empathetic understanding has benefited a wide range of downstream applications, such as medical dialogue systems [36], counseling conversation [104], and social chatbots [170]. The ability of empathy is to perceive, understand, and respond to another individual's feelings, experiences, and situations [99, 32]. It emphasizes understanding of the other. Empathy is composed of two aspects [30], which are (i) affection, i.e., emotion understanding and appropriate emotional reaction [53], and (ii) cognition, i.e., comprehension and reasoning of the other's experiences and situation [106]. Affection has been naturally considered in the empathetic conventional system in various ways [82, 89, 74]. Despite their success, their models lack understanding of the user's experiences, resulting in very weak empathy. The cognitive aspect of empathy should be involved, thereby enhancing users' comprehension. To this end, we propose to consider emotional causalities in conversations. On the one hand, emotional causality is an integral component of emotional experience [57]. On the other hand, it is recognized as an important feature in the analysis of human emotion [148]. In this part, two novel approaches are proposed, each employing a distinct reasoning mechanism to improve empathetic understanding. The first approach leverages external commonsense knowledge graphs to support reasoning over emotional causality, allowing the model to incorporate this broader knowledge. The second approach enables the model to infer emotional causalities internally, learning to identify and reason about the cause-and-effect relationships of emotions directly from conversations.

Chapter 3

Understanding with Multi-hop over Knowledge Graph

3.1 Introduction

This chapter introduces a novel method for enhancing empathetic understanding by leveraging emotional causality. Our approach aims to bridge the gap between a user’s emotions and their underlying experiences by extracting emotional causality paths from an external commonsense knowledge graph and incorporating them into response generation. The process involves three main steps. First, we identify emotion cues and cause cues from the conversation. Next, we employ multi-hop reasoning over a commonsense graph to infer paths from the causes to the emotions, constructing an emotional causality graph. These graphs are then encoded by multi-layer Graph Convolutional Networks (GCNs). The encoded graphs are integrated into the response generation in an implicit and explicit way. Implicitly, a fused, graph-aware representation is fed into the decoder to generate a generic vocabulary distribution for the empathetic response. Explicitly, to explicitly leverage the emotional causality information, we perform a second round of multi-hop reasoning directly on the causality graphs during each generation step. This process yields a vocabu-

lary distribution over the graph’s nodes, which is then used to refine the final output. We name this novel approach **Graph-based multi-hop Reasoning on Emotional Causality (GREC)**. The experimental results on the **EMPATHETICDIALOGUES** [109], a benchmark for empathetic dialogue generation, suggest that the proposed methods improve the model’s understanding of the user’s feelings and experiences. In addition, GREC is able to significantly improve the empathy level of generated responses.

To summarize, the main contributions of this chapter are as follows:

- (1) It integrates emotional causality reasoning into empathetic response generation, enabling a deeper and more nuanced understanding of the user’s emotional experiences.
- (2) It introduces a novel multi-hop reasoning mechanism over an external commonsense knowledge graph to model the formation process of user emotions. This inferred knowledge is then utilized to guide response generation.
- (3) It empirically verifies the effectiveness of the proposed model and evaluates the contribution of each component. Experimental results show that *GREC*, with emotional causality reasoning, significantly improves the empathetic quality of responses compared to baseline models.

3.2 Method

The overview architecture of the proposed model, *GREC*, is shown in Fig. 3.1. Based on a transformer model, emotional causalities are used to augment both the encoder and the decoder with a graph-based approach. For the encoder, we construct a series of emotional causality graphs for the conversation through

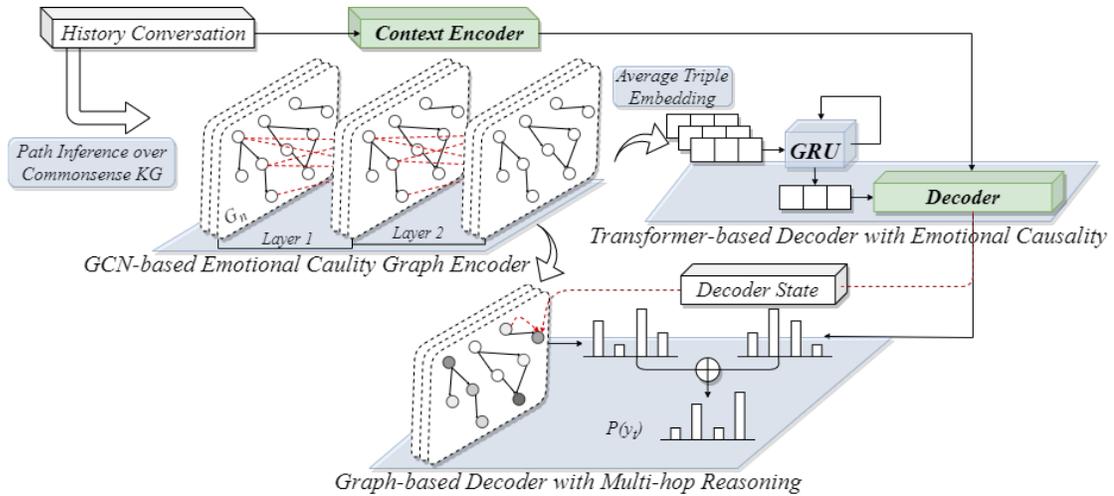


Figure 3.1: GREC: Overview architecture.

multi-hop reasoning over ConceptNet (Section 3.2.1), and then apply a multi-layer GCN encoder to model these emotional causality graphs (Line 19). Besides, we use a typical transformer encoder for context understanding (Section 3.2.2). For the decoder, we combine both the sequential information obtained by the transformed-based conversation context encoder and the structural knowledge obtained by the graph-based emotional causality encoder to compute the generic vocabulary distribution (Section 3.2.3). Additionally, a graph-based decoder with multi-hop reasoning computes the distribution over the concepts of the graph to enhance the generic vocabulary distribution for better generating the response (Section 3.2.3).

3.2.1 Emotional Causality Reasoning and Encoding

Emotional Causality Graph Construction

To understand the emotional causality of the user’s emotional experience in the conversation context, we construct a series of emotional causality graphs based on the given context. Specifically, we first identify the emotion concepts and

the cause concepts in the dialogue context and then reason out the relations between emotion concepts and cause concepts from the external commonsense knowledge graph, namely, ConceptNet, through a multi-hop strategy for constructing the emotional causality graphs.

To derive the emotion and cause concepts, we first locate the emotion and cause clauses within the dialogue context. Specifically, we treat each clause in the immediate utterance \mathcal{Q} as an emotion clause, since the objective is to generate an empathetic response to this utterance. Rather than constructing a single graph for the entire utterance, we build a separate emotional-causality graph for each clause of \mathcal{Q} , because different clauses may correspond to distinct causes. For every emotion clause, its potential cause clauses are identified from the conversation history using the RTHN model [152], which determines whether a clause serves as the emotional cause of another clause within a given context. These cause clauses are then concatenated to form the cause context for the corresponding emotion clause. Finally, we extract verbs, nouns, adjectives, and adverbs from each emotion clause and its associated cause clauses as the emotion concepts and cause concepts, respectively.

To construct the emotional causality graphs based on the emotion concepts and cause concepts, we adopt multi-hop reasoning over ConceptNet to infer semantic paths from emotion concepts to cause concepts, as shown in Fig. 3.2. It illustrates the process of deducing causal paths over an external commonsense graph starting from the cause concept *experience*. Underlined terms in red correspond to the cause concepts in the cause context, while underlined terms in blue correspond to the emotion concepts in the emotional clause. The *green space* represents a segment of the external commonsense knowledge graph. All candidate concepts are retained throughout the procedure to preserve rich seman-

tic information. We then perform multi-hop emotional causality reasoning on each clause pair that exhibits a cause–emotion relationship, iteratively deducing potential intermediate concepts from the complete set of cause concepts. The detailed steps for discovering these intermediate concepts and their relations, which help bridge cause and emotion concepts, are illustrated in Algorithm 1. Specifically, we maintain a concept stack S to store candidate concepts for inclusion in the emotional-causality graph. Initially, S is populated with all cause concepts in $C_{cans}^{(i)}$. At every hop h , we pop elements from S and gather their unvisited neighboring concepts (lines 6–10). From these neighbors, the top- K concepts—ranked by semantic similarity—are pushed back into S (lines 11–17), where semantic similarity is measured by the cosine similarity of their word embeddings. If an emotion concept v_e is encountered in S , it is popped immediately without further neighbor expansion. After H hops, we obtain the final emotional-causality graph, in which all intermediate concepts and edges are preserved, mimicking how humans recall and connect related ideas when composing a conversational response.

Graph-based Emotional Causality Encoder

Graph neural networks have been proven effective for producing graph-aware representations of nodes in the graph via integrating neighbors’ information of each node [169]. Therefore, to obtain the representation of concepts and relations in the emotional causality graphs, we apply multi-layer GCN [67] encoders to encode the emotional causality graphs. Moreover, following the idea of the TransE model [164], we update a concept embedding with the subtraction between each neighbor concept embedding and the corresponding relation embedding to obtain the relation representation.

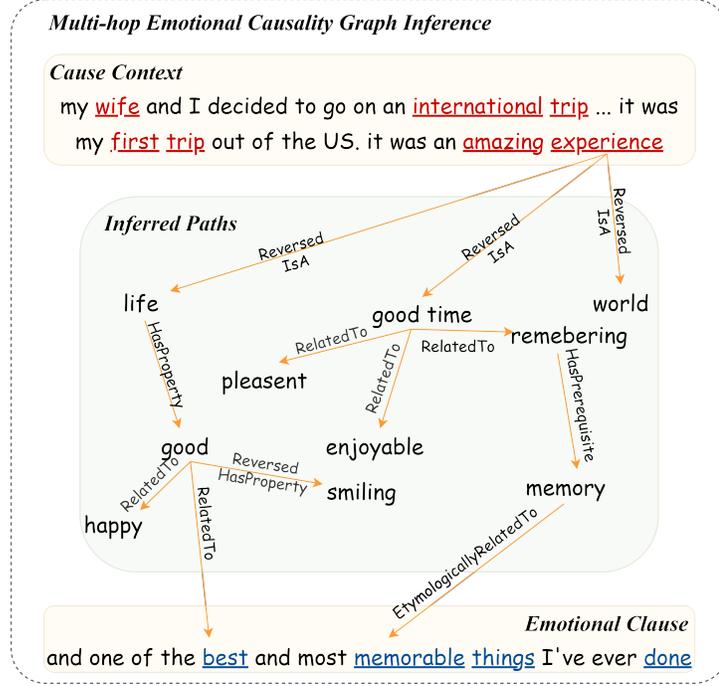


Figure 3.2: GREC: An example of emotional causality graph construction.

For each emotion clause, there is a constructed emotional causality graph $G = (V, R)$, where V denotes concepts and R denotes relations. The concepts in V are initialized by pretrained word embeddings, and the relations in R are initialized with randomly initialized relation-type embeddings. For each concept v_i , we update its embedding at the $(l + 1)^{th}$ layer by aggregating its neighbours \mathcal{N}_i including pairs of the concept and the relation linking to v_i :

$$h_i^{(l+1)} = \sigma \left(W_s^{(l)} h_i^{(l)} + \sum_{(j,r) \in \mathcal{N}_i} \frac{1}{|\mathcal{N}_i|} W_n^{(l)} (h_j^{(l)} - h_r^{(l)}) \right), \quad (3.1)$$

where $h_i^{(l)}$, $h_j^{(l)}$ and $h_r^{(l)}$ are the embeddings of node v_i , node v_j , and the relation between v_i and v_j at layer $(l)^{th}$; $W_s^{(l)}$ and $W_n^{(l)}$ are the two trainable parameter matrices specific to the layer l^{th} ; and σ is a non-linear active function. The relation embedding is also updated at the $(l + 1)^{th}$ layer via a linear active function:

$$h_r^{(l+1)} = W_R^{(l)} h_r^{(l)}. \quad (3.2)$$

Input: *ConceptNet*, G_C ;

The sets of cause concepts and emotion concepts:

$$V_{cas} = \{v_1^{cas}, \dots, v_M^{cas}\}, V_{emo} = \{v_1^{emo}, \dots, v_N^{emo}\}.$$

Output: An emotional causality graph, G .

```

1 Initialize an emotional causality graph  $G = (v_{cas}, \emptyset)$ ;
2 Initialize a concept stack  $S$  with cause concepts;
3 Initialize an empty temporal triple list  $T$ ;
4 for  $h$  from 1 to  $H$  do
5   while  $S$  is not empty do
6     pop  $v_i$  from  $S$ ;
7      $\mathcal{N}_i$ : the neighbouring concepts of  $v_i$  in ConceptNet;
8     foreach  $v_j$  in  $\mathcal{N}_i$  do
9       | append  $(v_i, r, v_j)$  into  $T$ ;
10    end
11    select the top  $K$  tail concepts which is most similar to the concepts
        in  $V_{emo}$  from  $T$ ;
12    foreach  $(v_i, r, v_j)$  in  $T$  do
13      | add  $v_j$  and  $r$  into  $G$ ;
14      | if  $v_j$  not in  $V_{emo}$  then
15        | push  $v_j$  into  $S$ 
16      | end
17    end
18  end
19 end

```

Algorithm 1: Framework of Emotional Causality Graph Construction by Multi-hop Reasoning over *ConceptNet*

After L iterations, the following contents can be obtained: $\{h_{v_1}^{(L)}, \dots, h_{v_{|V|}}^{(L)}\}$, a set of concept representations, and $\{h_{r_1}^{(L)}, \dots, h_{r_{|R|}}^{(L)}\}$, a set of relation representations.

3.2.2 Conversation Context Encoding

We utilize a transformer encoder [129, 33] for conversation context understanding. Same as previous works [82, 89, 74], we flatten all conversation context in C , and then add a special token *CLS* at the beginning of the input, denoting the

global memory of the whole sequence.

$$\bar{C} = [CLS; C]. \quad (3.3)$$

After that, the input tokens are converted to word embeddings. To embody the influence of the emotional cause and effect, we incorporate fixed causal embeddings into the input to show the importance of the clause in terms of their frequency as an emotional cause. In addition, we adopt dialogue state embeddings to distinguish situation text, the speaker’s utterance, and the listener’s utterance. The representation of the encoder input is as follows:

$$E^C = emb(\bar{C}) + emb_{cas}(\bar{C}) + emb_{utt}(\bar{C}). \quad (3.4)$$

Here, emb , emb_{cas} , and emb_{utt} represent the token embedding layer, causal embedding layer, and the dialogue state embedding layer, respectively. Finally, we use Transformer [129] to encode the input.

$$S^C = TRS_{enc}(E^C), \quad (3.5)$$

where $S^C \in \mathcal{R}^{L_c \times d_{model}}$, and L_c is the length of the input sequence.

To better learn the context representation, we introduce an auxiliary task of emotion prediction, which is also adopted in previous methods [82, 89, 74]. Specifically, we use the global memory of the whole sequence S_0^C to predict the emotion label q_e as follows:

$$q = S_0^C, \quad (3.6)$$

$$q_e \sim P(e|C) = \sigma(W_e q). \quad (3.7)$$

3.2.3 Emotional Causality Augmented Decoding

The emotional causality augmented decoder for empathetic response generation consists of two components, namely, a transformer-based decoder involv-

ing the overall emotional causality information and a graph-based decoder with multi-hop reasoning on the emotional causality graphs. These two decoders compute a generic vocabulary distribution and a graph-related concept distribution, respectively. And the final vocabulary distribution is a gated combination of these two vocabulary distributions. Under this setting, the influence of emotional causality can be strengthened through both implicit and explicit mechanisms.

Transformer-based Decoder with Emotional Causality

We integrate emotional causality with the transformer-based decoder by first using a GRU [25] to derive a holistic representation of emotional causality relations, which is subsequently fused into the generation process of the generic vocabulary distribution.

A sequence of encoded emotional causality graphs is constructed for every conversation. For each graph, its representation is computed through average pooling of the triple vectors $(v_i, r, v_j) \in \mathcal{G}_n$, giving

$$h_{\mathcal{G}_n} = \sum_{(v_i, r, v_j) \in \mathcal{G}_n} \frac{1}{|\mathcal{G}_n|} [h_i^{(L)}; h_r^{(L)}; h_j^{(L)}], \quad (3.8)$$

where $[;]$ denotes vector concatenation. Given the sequence of emotional causality graph representations $\{\mathcal{G}_1, \dots, \mathcal{G}_{N_q}\}$, we employ a bidirectional *Gated Recurrent Unit* (GRU) [25] to encode this sequence, which is

$$\begin{aligned} \overrightarrow{s}_n &= GRU(\overrightarrow{s}_{n-1}, h_{\mathcal{G}_n}), \\ \overleftarrow{s}_n &= GRU(\overleftarrow{s}_{n-1}, h_{\mathcal{G}_{N_q-n+1}}). \end{aligned} \quad (3.9)$$

Then, we concatenate the forward last hidden state and the backward last hidden state of the GRU cells to combine the forward and backward representations of the emotional causality:

$$s_{N_q} = [\overrightarrow{s}_{N_q}; \overleftarrow{s}_{N_q}]. \quad (3.10)$$

Finally, we obtain the emotional causality vector, representing the overall emotional causality, through a linear transformation:

$$\mathcal{H}_Q = \sigma(W_g s_{N_q}), \quad (3.11)$$

where W_g is the trainable parameter matrix.

Then, we integrate the emotional causality vector into the transformer decoder. Similar to the context encoder, we use a special token *SOS* together to represent the universal information of the decoder input $y_{<t}$. Then we append the emotional causality vector \mathcal{H}_Q to this special vector. That is

$$E_t = \text{emb}([y_{<t}]), \quad (3.12)$$

$$s_t = \text{trs}_{dec}([\text{emb}(\text{SOS}) + \mathcal{H}_Q; E_t], S^C). \quad (3.13)$$

Finally, we compute the generic vocabulary distribution:

$$P(\mathcal{V} | s_t, \mathcal{H}_Q) = \text{softmax}_{w \in \mathcal{V}}(W_{voc} s_t + b). \quad (3.14)$$

Graph-based Decoder with Multi-hop Reasoning

Inspired by the idea of *multi-hop reasoning flow* [59] and *pointer-generator decoder* [10], we adopt a graph-based decoder, which computes the concept distribution with multi-hop reasoning on the emotional causality graphs. At each decoding step, we adjust the weight of each concept in the graph by combining its neighbouring evidence and the current decoder state. Initially, scores of all cause concepts are set to 1, while other concepts are assigned a score of 0. Then, we broadcast information about scored concepts on the graph to update the unvisited concepts. For an unvisited concept $v_i \in V$, $\text{score}(v)$ is computed by aggregating the evidence from its visited neighbouring concepts \mathcal{N}_i^{in} :

$$\text{score}(v_i) = \sum_{(v_j, r) \in \mathcal{N}_i^{in}} \frac{1}{|\mathcal{N}_i^{in}|} (\gamma \cdot \text{score}(v_j) + R(v_j, r, v_i)). \quad (3.15)$$

Here, γ is a discount factor and $R(v_j, r, v_i)$ is the triple relevance under the current decoder output, which can be computed as below:

$$h_{(j,r,i)} = [h_j^{(L)}; h_r^{(L)}; h_i^{(L)}], \quad (3.16)$$

$$R(v_j, r, v_i) = \sigma(h_{(j,r,i)} W_{rel} s_t). \quad (3.17)$$

After L -hop interactions, the distribution over the concepts is as follows:

$$P(V | s_t, \mathcal{G}) = \text{softmax}_{v \in V}(\text{score}(v)). \quad (3.18)$$

The final generation distribution conjoins the distribution over the concepts in the emotional causality graphs V and the distribution over the standard vocabulary \mathcal{V} with a soft gate g_t , which decides whether or not to refer to the emotional causality graphs:

$$g_t = \sigma(W_g s_t), \quad (3.19)$$

$$o_t \sim P(y_t | y_{<t}, C, \mathcal{G}, \mathcal{H}_Q) = g_t \cdot P(V | s_t, \mathcal{G}) + (1 - g_t) \cdot P(\mathcal{V} | s_t, \mathcal{H}_Q). \quad (3.20)$$

3.2.4 Training Objectives

To enhance the empathy of the model, the model is trained via multi-task learning with an auxiliary task to predict the emotion label of the speaker. We compute the multi-task loss with the prediction results $q_e \sim P(e|C)$, which is computed in Equation (3.7).

$$L^e = - \sum_{e=1}^{|E|} p_e \log(q_e), \quad (3.21)$$

where p_e is the real label probability of the speaker's emotion (1 for the true emotion label, and 0 for the others), and q_e is the predicted one (Section 3.2.2).

The loss of the response generation task is:

$$L^g = - \sum_{t=1}^{|Y|} \sum_{v=1}^{|V|} p_t^v \log(o_t^v), \quad (3.22)$$

where Y is the target output, p_t^v is the true probability of the token w at the position t in Y , and o_t^v is the predicted one.

The total loss of the model is a weighted sum of these two components through a coefficient λ :

$$Loss = L^g + \lambda L^e. \quad (3.23)$$

3.3 Experiments

3.3.1 Experimental Settings

Dataset Experiments were conducted on the popular benchmark EMPATHETICDIALOGUES [109]. Table 3.1 shows an example from EMPATHETICDIALOGUES. Both methods were tested on this dataset. The dataset consists of 25k crowd-sourced one-on-one conversations. Each conversation is developed between a speaker and a listener, along with a label indicating the speaker’s emotion and a paragraph describing their emotional situation. There are 32 emotion categories distributed in a balanced way. The speakers talk about their situations, and the listeners attempt to understand the speakers’ feelings and reply accordingly. At the training time, the emotional labels of the conversations are given in terms of the degree of empathy to facilitate multi-task learning, while the labels are not provided when testing the model and evaluating the generation performance. Following its original division, we adopt approximately 80%, 10%, and 10% of the dataset for training, validation, and testing.

Comparison Models The following seven models are selected for comparison in the task of empathetic response generation:

- **Transformer with Multitask (Multi-TRS)**. This model employs a univer-

Table 3.1: GREC: A conversation from EmpatheticDialogue.

Label: Excited
Situation: My wife and I took a 3 week trip to New Zealand last year.
Conversation:
<i>Speaker:</i> My wife and I decided to go on an international trip last year, and it was my first trip out of the US. It was an amazing experience, and one of the best and most memorable things I've ever done.
<i>Listener:</i> That's awesome! How helpful were the other people so that you could enjoy your trip?
<i>Speaker:</i> My wife was amazing. She booked the whole trip for us, and planned out some things for us to do while we were there.
<i>Listener:</i> Must have been super fun.

sal transformer encoder-decoder architecture [129] and is trained under a multi-task learning paradigm. The overall loss is computed according to Eq. (3.21)–Eq. (3.23).

- **MoEL.** The Mixture of Empathetic Listeners (MoEL) [82] consists of several emotion-specific decoders along with a meta decoder. Each emotion-specific decoder is tailored to respond to a particular emotion, while the meta decoder combines outputs from all decoders to produce the final empathetic response. Our implementation follows the original parameter settings and reported values in [82].
- **MIME.** Building upon MoEL, MIME [89] incorporates polarity-based emotion clusters and emotional mimicry to generate more balanced empathetic responses. It introduces stochastic elements in the emotion mixture, allowing the model to produce diverse responses that align with the emotional context.
- **EmpDG.** This model [74] leverages both dialogue-level and token-level emotional information to capture fine-grained nuances in the speaker's emotions. Additionally, it incorporates speaker feedback within an inter-

active adversarial framework to enhance the expression of empathy.

All the above models consider merely the users' emotions.

Implementation Details *Pytorch* is used in the model implementation. The token embedding is initialized with pre-trained Glove vectors of dimension 300 [103] and shared by the transformer embedding layer, the projection layer of the decoder, and the graph-based encoder.

For the transformer encoder–decoder, we employ 6 layers with 8 attention heads. The query, key, and value embeddings each have a dimensionality of 40. The position-wise feed-forward sublayer is replaced by a 1D convolutional block with 50 filters of width 3. Clause-level causal embeddings, used to indicate the importance of clauses in the conversation context, are randomly initialized and kept fixed during training, with the dimensionality empirically set to 300. For emotion–cause extraction, we follow the original RTHN approach [152] and pretrain an emotion-causality detection model on the REMAN dataset [62]. The resulting RTHN achieves an F1 score of 0.95, showing strong accuracy in identifying emotion causes. Optimization uses Adam [65] with a batch size of 32 and an initial learning rate of 0.1. The learning rate decays by a factor of 0.1 every 500 steps until reaching 1×10^{-5} . Beam search is performed with a beam size of 5. For statistical significance testing, the model is trained for 10 epochs.

3.3.2 Evaluation

Automatic Evaluation Metrics Three kinds of metrics are applied for automatic evaluation: (1) Perplexity (PPL) [5], which measures the model's confidence in the response generation. (2) BLEU [100], which estimates the matching between n-grams of the generated response and those of the golden response.

Model	PPL↓	BLEU-3↑	BLEU-4↑	P _{BERT} ↑	R _{BERT} ↑	F _{BERT} ↑
MoEL	36.87±0.97	4.53±0.53	2.80±0.32	.499±.008	.467±.007	.480±.006
MIME	37.88±0.49	4.48±0.15	2.71±0.09	.490±.004	.466±.002	.475±.002
EmpDG	55.64±3.78	3.64±0.38	1.99±0.22	.475±.007	.458±.008	.465±.004
GREC	32.66* ±0.32	4.96* ±0.14	3.16* ±0.08	.532* ±.003	.506* ±.003	.512* ±.004

Table 3.2: GREC: Automatic evaluation results.

BLEU-3 and **BLEU-4** are adopted. (3) BERTScore [160], which computes the similarity for each token in the generated response with that in the golden response. We use its matching precision, recall and F1 score (P_{BERT} , R_{BERT} , and F_{BERT}). For perplexity, a lower score indicates a better performance, while for the rest of the metrics, higher scores indicate better performance. Table 3.2 demonstrates the automatic results. For each method, multiple runs are repeated with different seeds. The average values of the results are displayed along with the standard deviations. The values marked with * mean the results are statistically significant with $p < 0.05$. The highest score in terms of each metric is in bold. It can be observed that the proposed model, GREC, outperforms other models in most of the evaluation metrics. The lowest perplexity score suggests that GREC is more confident in its generated responses than other models. The highest BLEU and BERTScore scores indicate that the approach can generate more human-like responses by incorporating causality reasoning. All the above advantages are especially significant and stable, which is evident in the high degrees of statistical significance and small standard deviations, respectively.

Human Ratings For qualitative evaluation, we conduct human assessments to compare GREC with other baseline models. Three annotators with diverse demographic backgrounds, including age, occupation, and nationality, rated the responses. We randomly selected 120 samples from the test set and evaluated them on three dimensions: **empathy** (the extent to which the response

Model	Empathy \uparrow	Relevance \uparrow	Fluency \uparrow
MoEL	2.73	2.63	4.82
MIME	2.30	2.24	4.88
EmpDG	2.31	2.27	4.52
GREC	2.79	2.87	4.84

Table 3.3: GREC: Human rating results.

reflects understanding of the speaker’s feelings), **relevance** (how appropriately the response aligns with the conversation context), and **fluency** (the naturalness and grammatical correctness of the response). Each dimension is scored on a 1–5 scale, where 1 indicates the lowest quality and 5 the highest. As shown in Table 3.3, differences in **fluency** across models are minimal, whereas *GREC* consistently achieves the top scores for **empathy** and **relevance**. We attribute this improvement to the integration of emotional causality. Unlike MoEL, MIME, and EmpDG, which rely primarily on emotion labels or keywords to represent the speaker’s emotions, emotional causality allows the model to capture how these emotions arise. By encoding rich semantic relations between causes and emotions and applying multi-hop reasoning over the emotional causality graph, *GREC* can focus on the most relevant information, resulting in responses that are both more empathetic and contextually appropriate.

Human A/B Test To evaluate the overall performance of various models, we conduct a human A/B comparison. For two models, A and B, in our study, *GREC* vs. *Multi-TRS*, *MoEL*, *MIME*, annotators are asked to determine which model produces the better responses. Annotators may choose *Tie* if the quality of responses from both models is similar, whether good or poor. As shown in Table 3.4, responses generated by *GREC* and *CREA* are generally preferred by human judges over those from *Multi-TRS*, *MoEL*, *MIME*, and *EmpDG*. This suggests that our proposed models are more capable of producing responses

GREC vs.	GREC wins	GREC loses	Tie
Multi-TRS	50.0%	22.5%	27.5%
MoEL	48.3%	20.8%	30.8%
MIME	35.0%	27.5%	37.5%
EmpDG	62.7%	11.7%	25.6%

Table 3.4: GREC: Human A/B test results.

that are both empathetic and contextually relevant.

According to Table 3.4, responses generated by *GREC* are more often preferred by human judges in general than the responses from *Multi-TRS*, *MoEL*, and *MIME*. This also indicates that our model is able to generate more empathetic and more relevant responses.

3.3.3 Further Analyses

Ablation Study To verify the effect and the contribution of each component of the proposed models, an ablation study is conducted. The following three settings are considered:

- (1) **w/o graph.** In this variant, the explicit graph representation of emotional causality is removed, and the causality information is provided as plain text. As a result, the emotional causality construction described in section Section 3.2.1 is omitted. For understanding emotional causality, we extract a sequence of clauses corresponding to emotional causes from the conversation context. These sequences are encoded using RNNs, replacing the GCN encoder introduced in section line 19. During decoding, both implicit and explicit forms of emotional causality are still leveraged. Instead of calculating the concept distribution as in Equation (3.18), we determine the weights of the emotion and cause concepts \mathcal{V} by

$$P(V | s_t, \mathcal{V}) = \text{softmax}_{w \in \mathcal{V}} (\text{emb}(w)s_t + b). \quad (3.24)$$

Method	PPL↓	BLEU-3↑	BLEU-4↑	P _{BERT} ↑	R _{BERT} ↑	F _{BERT} ↑
<i>w/o graph</i>	33.93	3.47	2.20	.485	.461	.473
<i>w/o implicit decoding</i>	33.50	4.90	3.05	.517	.475	.496
<i>w/o explicit reasoning</i>	32.94	4.62	2.81	.500	.463	.477

Table 3.5: GREC: Automatic evaluation results of ablation study.

Method	Empathy↑	Relevance↑	Fluency↑
<i>w/o graph</i>	2.54	2.54	4.80
<i>w/o implicit decoding</i>	2.72	2.73	4.70
<i>w/o explicit reasoning</i>	2.62	2.62	4.93

Table 3.6: GREC: Human rating results of ablation study.

- (2) **w/o implicit decoding.** To study the effect of implicit emotional causality, we remove its integration within the transformer decoder as described in section Section 3.2.3 (denoted as *w/o encoder*). Consequently, Equation (3.13) is replaced with

$$s_t = trs_{dec}([emb(SOS); E_t], S^C). \quad (3.25)$$

- (3) **w/o explicit reasoning.** We ablate the explicit use of the emotional causality by removing the graph-based decoder with multi-hop reasoning in section Section 3.2.3. The final vocabulary distribution is calculated by Equation (3.14) only.

The results in Table 3.5 and Table 3.6 show that all components contribute to GREC because removing any of them decreases the overall performance. Moreover, each component contributes to GREC from different aspects. Based on the observation and the design of the model, we have the following analysis:

- (1) Emotional causality graphs contribute the most to the model performance. When the graph structure is removed, performance in both empathy and relevance drops significantly by 9.03% and 11.45%, respectively. These graphs, enriched with external commonsense knowledge, provide a more

robust representation of emotional causality. While leveraging the emotional cause clause alone improves performance over baselines such as Multi-TRS and MoEL in empathy and relevance, emotional causality graphs offer a more semantically rich relationship between cause and emotion. Moreover, the graphs support multi-hop reasoning, supplying the decoder with more informative signals.

- (2) Decoding with the emotional causality in an implicit way also benefits the model. Compared to the other two ablated variants, it exhibits smaller performance degradation across most evaluation metrics. However, fluency sees a notable decline. This may stem from the implicit nature of emotional causality integration within the transformer decoder—while this approach does not produce dramatic performance gains, it allows emotional causality to subtly shape the vocabulary distribution, making its influence more organic.
- (3) Graph-based decoding with multi-hop reasoning further enhances the model, especially in empathy and relevance. This component enables the decoder to reason over the emotional causality graph, extracting key information that supports more contextually appropriate and emotionally resonant responses. Nevertheless, it negatively affects fluency, likely because emotional causality graphs include not only helpful cues but also noisy information. Future work could focus on refining graph construction or selectively filtering information to reduce noise and improve response fluency.

Case Study Table 3.7 shows two cases generated by *GREC*. These utterances are from continuous dialogues in the test set. The gold responses are also displayed. Terms reflecting empathy are highlighted in *green*. The results indicate that *GREC* more accurately captures the speaker’s emotions and gener-

ates responses that are natural and contextually appropriate compared to other models. In both example conversations, *GREC* produces responses that convey suitable emotions and expressions. For instance, it employs verbs such as “get it back” and “complain” when reacting to a negative attitude, demonstrating specificity and a nuanced understanding of the speaker’s feelings. By contrast, other models tend to rely on generic emotional words, which may sound less natural and could apply to a variety of situations. *GREC* also effectively identifies key concepts driving the speaker’s emotions. In the second conversation, it references “shadow,” which is central to that dialogue. Moreover, *GREC* generally shows improved performance from the second turn onward, a trend observed in these examples. These observations suggest that emotional causality graphs encode critical contextual information and enable the model to leverage it for generating more appropriate responses at each turn.

To investigate how emotional causality information contributes to empathetic response generation, we visualize the attention scores of concepts within the emotional causality graph during sentence generation. Fig. 3.3 presents an example graph for the speaker’s second utterance, “*It was so scary ... an intruder!*”, showing how attention weights for different concepts evolve across the decoding steps. Here, *red circles* denote the cause concepts, *blue circles* indicate the emotion concepts, and *green circles* mark the concepts that appear in the final generated response. The area of each circle is proportional to its attention weight at the corresponding decoding step.

The related concepts, which are suitable for this conversation, are inferred and selected. Moreover, the attention weights of the concepts varying over the decoding steps are reflected in this figure. The attention weights for the trivial terms, e.g., “war” and “storm,” decrease, while weights for necessary terms,

Conversation 1	
Speaker Emotion: DISGUSTED	
Situation: I once went to a pizza place and found a hair on my pizza. I was grossed out.	
Speaker	I had pizza at a restaurant a couple days ago and found a hair on my pizza!
Gold	It is <i>disgusting</i> . What did you do?
Transformer	What did you do ?
GREC	That is so <i>gross</i> ! Did you get it back?
Speaker	I told the waiter and demanded a refund! It was gross.
Gold	You did good. You should have called the manager to tell him know.
Transformer	Wow, that is <i>gross</i> . I hope you were able to fix it .
GREC	It is so <i>gross</i> . Did you <i>complain</i> to the manager ?
Conversation 2	
Label: TERRIFIED	
Situation: One night my children and I came home, and I saw a shadow inside the home when I came inside. My heart jumped into my throat and you could hear my scream in the next town for fear!	
Speaker	One night when my children and I came home, I saw a shadow inside! I was so fearful and afraid!
Gold	Oh that can be <i>scary</i> . I don't watch horror movies because I know that I'll do just that!
Transformer	<i>Terrified</i> , oh no! I am so sorry to hear that.
GREC	Oh no! I <i>hate</i> when that happens to me. What was it?
Speaker	It was so scary! Luckily, it was actually my shadow! The light made my shadow show up, but it first appeared to be an intruder!
Gold	Well I'm glad you were ok! Sometimes it's not just a shadow.
Transformer	<i>Terrified</i> , I am sure it was a <i>scary movie</i> .
GREC	I can imagine. I would have been <i>terrified</i> .

Table 3.7: GREC: Case study.

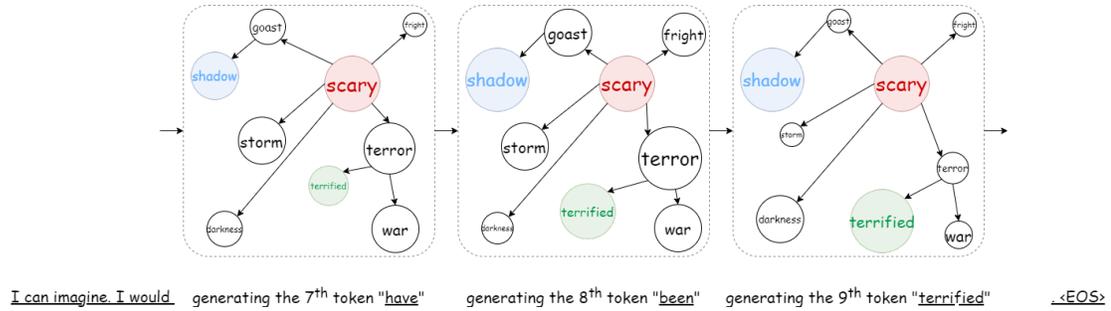


Figure 3.3: GREC: A case of the emotional causality graph.

such as “terrified” and “darkness,” increase over steps. In addition, “terrified” is even selected in the response. This indicates that the model can reason out important information from the emotional causality graphs for empathetic response generation through a multi-hop strategy. Additionally, concepts with higher scores are deduced from the cause concepts. This significantly enhances model transparency and explainability.

3.4 Conclusion

This chapter explores the involvement of emotional causalities to deepen the empathetic understanding of the user’s states in response generation. Specifically, a method named GREC is proposed to incorporate emotional causalities in empathetic response generation. These emotional causalities are extracted with a multi-hop reasoning over an external knowledge graph. Through the reasoning process, the multi-hop explores how the emotion emerges and provides a deep understanding of the user’s emotion. With a typical transformer model as the skeleton, we augment the encoder and decoder with emotional causality. Experiments were conducted on the benchmark, EmpatheticDialogues. Results prove that *GREC* outperforms competitive models, including those state-of-the-art. We show that graph-based reasoning provides better model transparency

and explainability in the ablation study and the case study.

Chapter 4

Emotional Causalities Generation for Understanding

4.1 Introduction

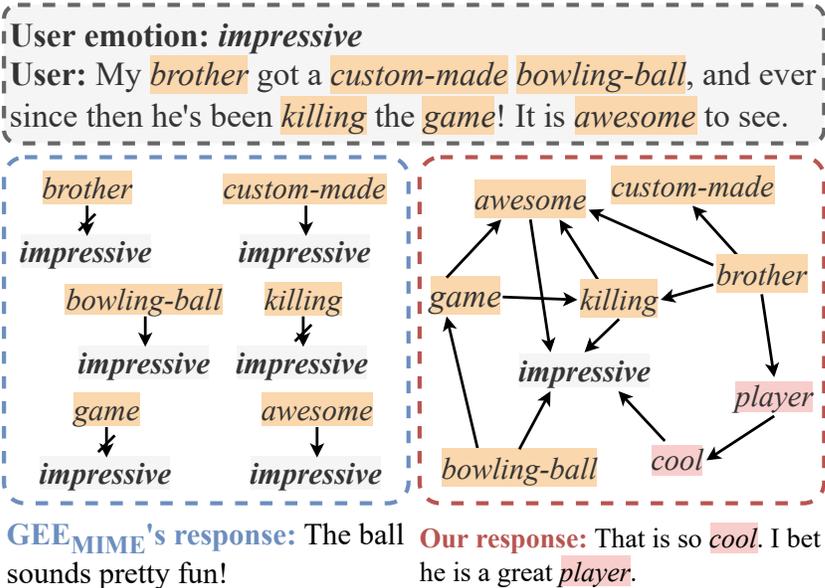


Figure 4.1: CARE: Causality reasoning for empathetic understanding.

The previous chapter demonstrated that incorporating emotional causalities can significantly enhance a model’s empathetic understanding of users. To obtain the emotional causalities, it adopts multi-hop reasoning over an exter-

nal commonsense knowledge graph. This chapter proposes an alternative approach to reasoning about emotional causalities by leveraging the model’s internal structure, thereby eliminating the reliance on external knowledge sources. Specifically, the model is named as **CARE (CAusality Reasoning for Empathetic conversation)**. CARE reasons all plausible causalities, i.e., causalities stated explicitly in the dialogue history and probably in the future dialogue, interdependently and simultaneously by formulating the reasoning as a conditional graph generation task. We propose a method to achieve a deeper understanding of a user’s feelings and experiences by generating a causal graph. In this graph, each node represents a word related to the user’s emotions or experiences, and each edge indicates a causal relationship between them. This graph is conditioned on the user’s emotion, the conversation’s history, and predicted future dialogue. To accomplish this, we design a Conditional Variational Graph Auto-Encoder (CVGAE), which is a conditional extension of the Variational Graph Auto-Encoder (VGAE) [66]. The CVGAE uses latent variables to perform conditional structure prediction, enabling it to model all plausible causalities. This approach allows the model to gain a deeper, more nuanced understanding of the causal factors behind a user’s emotions and experiences. In addition, certain feelings and experiences—though not explicitly mentioned in the dialogue history—can be inferred during this process, as illustrated in Fig. 4.1. This implicit understanding plays a crucial role in generating empathetic responses, which previous methods are unable to achieve. Then, CARE infuses all plausible causalities into response generation using a multi-source attention mechanism in the decoder. In addition, we adopt multi-task learning to integrate causality reasoning and response generation during training. The experimental results on the benchmark EMPATHETICDIALOGUES [109] prove the effectiveness of CARE.

In summary, the primary contributions of this chapter are outlined as follows:

- (1). We introduce an approach to reason over all potential causalities in empathetic dialogues in an interdependent and simultaneous manner, enabling a deeper comprehension of the user’s emotions and experiences through the internal model structure.
- (2). Causality reasoning is formulated as a conditional graph generation problem, and we employ CVGAE, which leverages latent variables for conditional structure prediction, to perform this reasoning effectively.
- (3). We present CARE, a model that enhances empathetic response generation by incorporating causality reasoning, and demonstrate its superior performance on the EMPATHETICDIALOGUES benchmark.

4.2 Method

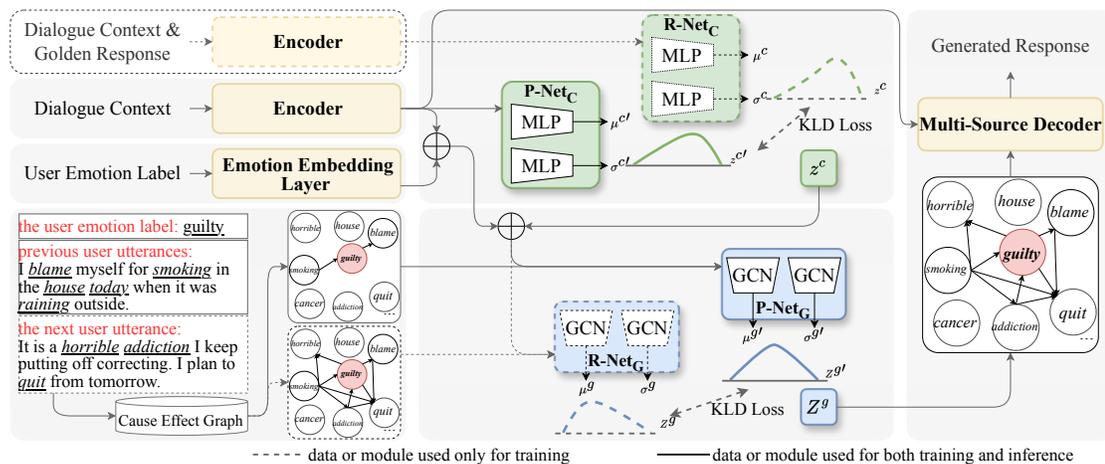


Figure 4.2: CARE: Overview architecture..

Fig. 4.2 illustrates the overall architecture of the proposed CARE model. Solid lines denote modules or data utilized for both posterior and prior com-

putations, whereas dashed lines indicate components used solely for posterior computations. CARE begins by reasoning over all potential causalities simultaneously through the construction of a causal graph. Specifically, CVGAE is employed to generate this graph using the user’s emotion, dialogue history, and predicted future dialogue content. It is important to note that CVGAE operates differently during training and inference: during training, it reconstructs a posterior causal graph (via $R\text{-Net}_G$) using the posterior graph as input, whereas during inference, it produces a posterior causal graph (via $P\text{-Net}_G$) based on a prior causal graph. The prior causal graph encodes causal relations explicitly present in previous user utterances, while the posterior graph incorporates additional causalities emerging in the next user utterance. Finally, CARE integrates the causalities from the reasoned graph into response generation through multi-source attention in the decoder.

4.2.1 Graph Construction

I need a prior causal graph $\mathcal{G}_{prior} = (\mathcal{V}, \mathcal{E}_{prior})$ and a posterior causal graph $\mathcal{G}_{post} = (\mathcal{V}, \mathcal{E}_{post})$, for causality reasoning at inference and training stage, respectively. We construct them with the assistance of a causal knowledge graph, i.e., Cause Effect Graph (CEG) [79]. These two graphs share the same node set, which theoretically contains all nodes in CEG. However, for effectiveness, we only consider those among a certain set of nodes \mathcal{V} , which contains emotion label words, words appearing in previous user utterances, and one-hop neighbors of the above two kinds of words. The edge sets of these two graphs are different. \mathcal{E}_{prior} contains causal relationships in previous user utterances, while \mathcal{E}_{post} also contains those in the next utterances. In specific, we collect \mathcal{E}_{prior} and \mathcal{E}_{post} according to following rules. For any couple nodes in \mathcal{V} having a relationship

in CEG, if both nodes are covered by the user emotion label word and words in previous user utterances, we add the relationship into \mathcal{E}_{prior} ; if both nodes are covered by the user emotion label word and words in previous and next user utterances, we add the relationship into \mathcal{E}_{post} .

4.2.2 Conditional Variational Graph Auto-Encoder

A novel structure CVGAE is designed to generate a (posterior) causal graph for causality reasoning. As an extension of VGAE, CVGAE works in a similar manner. In particular, it generates graphs' latent variables for graph reconstruction under some conditions, including a context condition, an emotion condition, and a context latent variable.

Context and Emotion Conditions The context condition is expected to provide information about the dialogue context C , thus it is derived from the encoder output. Following [140], we use multi-head attention to perform it. That is:

$$\mathbf{c}_{\text{ctx}} = \text{MultiHead}(v_{\text{rand}}, E_{\text{out}}, E_{\text{out}}), \quad (4.1)$$

where $E_{\text{out}} \in \mathbb{R}^{|C| \times d}$ is the encoder output computed by $\text{TRS}_{\text{enc}}(C)$ and d is the hidden size; $v_{\text{rand}} \in \mathbb{R}^{1 \times d}$ is a randomly initialized vector and is regarded as a single query for multi-head attention.

The emotion condition is expected to provide information about the user's emotion e . Accordingly, we define the emotion embedding $E^{\text{emo}} \in \mathbb{R}^d$, which converts an emotion label into embeddings. The emotion condition is formulated as:

$$\mathbf{c}_{\text{emo}} = E^{\text{emo}}(e). \quad (4.2)$$

Context Latent Variable I use a context latent variable z_c to provide information from the future dialogue. This variable is generated by a contextual recognition net ($R\text{-Net}_C$ in Fig. 4.2) with dialogue context C and the golden response R as input:

$$q_c(\mathbf{z}^c|C, R) = \mathcal{N}(\mathbf{z}^c|\mu^c, \sigma^{c^2}). \quad (4.3)$$

Here $\mu^c = \text{MLP}_\mu(\mathbf{c}_{\text{lant}})$ is the mean vector and $\log\sigma_c = \text{MLP}_\sigma(\mathbf{c}_{\text{lant}})$ is the log-variance vector, where \mathbf{c}_{lant} is accessed similar to Eq. (4.1):

$$E_{rep} = \text{TRS}_{\text{enc}}(C \oplus R), \quad (4.4)$$

$$\mathbf{c}_{\text{lant}} = \text{MultiHead}(v_{\text{rand}}, E_{rep}, E_{rep}). \quad (4.5)$$

Graph Latent Variables We generate graph latent variables \mathbf{Z}^g by a recognition net ($R\text{-Net}_G$ in Fig. 4.2): $q_g(\mathbf{Z}^g|\mathcal{V}, \mathbf{A}_{post}, c_{\text{cond}})$, where \mathbf{A}_{post} is the adjacency matrix of \mathcal{G}_{post} .

$$q(\mathbf{z}_i^g|\mathcal{V}, \mathbf{A}_{post}, c_{\text{cond}}) = \mathcal{N}(\mathbf{z}_i^g|\mu_i^g, \boldsymbol{\sigma}_i^g),$$

$$\text{with } \mu^g = \text{GCNLayer}_\mu(\mathcal{H}_\mathcal{V}, \mathbf{A}_{post}), \quad (4.6)$$

$$\text{and } \log\sigma^g = \text{GCNLayer}_\sigma(\mathcal{H}_\mathcal{V}, \mathbf{A}_{post}).$$

Here, \mathcal{N} is a sampling function following the Gaussian distribution. μ is the matrix of the mean vectors μ_i ; $\log\sigma$ is the matrix of log-variance vectors $\log\sigma_i$. The shared hidden state $\mathcal{H}_\mathcal{V}$ is generated with attention to the concatenation of \mathbf{c}_{ctx} , \mathbf{c}_{emo} , and \mathbf{z}^c :

$$c_{\text{cond}} = \mathbf{c}_{\text{ctx}} \oplus \mathbf{c}_{\text{emo}} \oplus \mathbf{z}^c,$$

$$\hat{\mathcal{H}}_\mathcal{V} = \text{GCNLayer}_h(\mathcal{V}, \mathbf{A}_{post}), \quad (4.7)$$

$$\mathcal{H}_\mathcal{V} = \text{MultiHead}(\hat{\mathcal{H}}_\mathcal{V}, c_{\text{cond}}, c_{\text{cond}}).$$

Causal Relation Generation With graph latent variables \mathbf{Z}^g , we reconstruct the posterior causal graph. The graph is represented by a matrix adjacency $\hat{\mathbf{A}}$,

which can be computed by:

$$p(\hat{\mathbf{A}}|\mathbf{Z}) = \prod_{i=1}^{|\mathcal{V}|} \prod_{j=1}^{|\mathcal{V}|} p(\hat{\mathbf{A}}_{ij}|\mathbf{z}_i, \mathbf{z}_j), \quad (4.8)$$

$$\text{with } p(\hat{\mathbf{A}}_{ij} = 1|\mathbf{z}_i, \mathbf{z}_j) = \text{sigmoid}(\mathbf{z}_i^\top \mathbf{z}_j), \mathbf{z}_i, \mathbf{z}_j \in \mathbf{Z}^g.$$

Then we select top- k relationships from the reconstructed graph according to their probability, denoted as $\mathcal{R} = (r_1, \dots, r_k)$, where r_i is the sum of the head and tail node embeddings.

Inference Stage During inference, R (the golden response) and \mathbf{A}_{post} are unavailable, thus we use a prior net $p_g(\mathbf{Z}^g|\mathcal{V}, \mathbf{A}_{prior}, c'_{cond})$ ($P\text{-Net}_G$ in Fig. 4.2) to approach $q_g(\mathbf{Z}^g)$, i.e., Eqs. (4.6) and (4.7). \mathbf{A}_{prior} is \mathcal{G}_{prior} 's adjacency matrix, and $c'_{cond} = \mathbf{c}_{ctx} \oplus \mathbf{c}_{emo} \oplus \mathbf{z}^{c'}$, where $\mathbf{z}^{c'}$ is obtained by a contextual prior net ($P\text{-Net}_C$ in Fig. 4.2):

$$p_c(\mathbf{z}^{c'}|C) = \mathcal{N}(\mathbf{z}^{c'}|\mu^{c'}, \sigma^{c'^2}), \quad (4.9)$$

with $\mu^{c'} = \text{MLP}_{\mu'}(\mathbf{c}_{ctx})$, $\log\sigma^{c'} = \text{MLP}_{\sigma'}(\mathbf{c}_{ctx})$.

4.2.3 Graph-Infused Response Generation

To infuse the reasoned \mathcal{R} into generation, we enable the decoder to attend to both dialogue context and the causal graph ($Multi\text{-Source Decoder}$ in Fig. 4.2). Typically, the original decoder includes L_{dec} decoder layers, each consisting of three sub-layers. The first one, i.e., the self-attention sub-layer, computes a representation of the input sequence:

$$\begin{aligned} \hat{H} &= \text{MultiHead}(H_{in}, H_{in}, H_{in}), \\ H_{out}^{(self)} &= \text{LayerNorm}(\hat{H} + H_{in}), \end{aligned} \quad (4.10)$$

where H_{in} is the embedding right-shifted response for the first decoder layer, and is the output of the $(l-1)$ -th decoder layer for the l -th decoder layer. Then

the decoder attends to the dialogue context by a cross-attention sub-layer:

$$\begin{aligned} H_{in} &= H_{out}^{(self)}, \\ \hat{H} &= \text{MultiHead}(H_{in}, E_{out}, E_{out}), \\ H_{out}^{(cross)} &= \text{LayerNorm}(\hat{H} + H_{in}). \end{aligned} \quad (4.11)$$

The output of the l -th decoder layer is obtained by the feed-forward sub-layer:

$$\begin{aligned} H_{in} &= H_{out}^{(cross)}, \\ H_{out}^{(ffn)} &= \text{LayerNorm}(\text{FFN}(H_{in}) + H_{in}). \end{aligned} \quad (4.12)$$

Finally, we apply linear transformation and a softmax operation on the output of the L_{dec} decoder layer to predict the token probability distribution at each token position t :

$$P_t = \text{softmax}(H_{out,t}^L W_o + b_o), \quad (4.13)$$

where $H_{out,t}^L$ is the final output for the t -th token; $W_o \in \mathbb{R}^{d \times d_{vocab}}$ and $b_o \in \mathbb{R}^{d_{vocab}}$ are parameters, and d_{vocab} is the vocabulary size.

In particular, we slightly modify the cross-attention sub-layer of the original decoder with the multi-source attention mechanism. The output after this modified sub-layer is computed by:

$$\begin{aligned} \hat{H}^C &= \text{MultiHead}(H_{in}^{(cross)}, E_{out}, E_{out}), \\ \hat{H}^R &= \text{MultiHead}(H_{in}^{(cross)}, \mathcal{R}, \mathcal{R}), \\ \hat{H} &= (\hat{H}^C \oplus \hat{H}^R) W_{multi}, \\ H_{out}^{(cross)} &= \text{LayerNorm}(\hat{H} + H_{in}^{(cross)}), \end{aligned} \quad (4.14)$$

where $E_{out} \in \mathbb{R}^{|C| \times d}$ is the encoder output, $W_{multi} \in \mathbb{R}^{2d \times d}$ is a group of linear transformation parameters, and H_{in} is the output of the self-attention sub-layer of the decoder computed by Eq. (4.10). Notably, the reset of the original decoder, i.e, Eqs. (4.10), (4.12) and (4.13), remains the same. In this way, we generate the final response.

4.2.4 Training Objectives

The model is optimized with multi-task learning to further integrate the causality reasoning and the graph-infused response generation. For the causality reasoning, we consider graph reconstruction accuracy and similarity between the posterior and prior distributions. The corresponding loss can be calculated by:

$$\begin{aligned} \mathcal{L}_r = & \mathbb{E}_{q_g(\mathbf{Z}^g | \mathcal{V}, \mathbf{A}_{post}, c_{cond})} [\log p(\hat{\mathbf{A}} | \mathbf{Z}^g)] \\ & - \text{KL}[q_g(\mathbf{Z}^g) || p_g(\mathbf{Z}^{g'})] \\ & - \text{KL}[q_c(\mathbf{z}^c) || p_c(\mathbf{z}^{c'})], \end{aligned} \quad (4.15)$$

where $\text{KL}[q(\cdot) | p(\cdot)]$ is the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$. The response generation loss is calculated by:

$$\mathcal{L}_g = \prod_{t=1}^{|R|} P_t, \quad (4.16)$$

where P_t is obtained by Eq. (4.13). Finally, we train CARE by maximizing $(\mathcal{L}_r + \mathcal{L}_g)$.

4.3 Experiments

4.3.1 Experimental Settings

Dataset Experiments were conducted on the popular benchmark EMPATHETICDIALOGUES[109]. Its original division 8:1:1 for training, validation, and testing is adopted in the experiments.

Comparison Models Seven models are selected for comparison, and they are: Transformer with Multitask (Multi-TRS), MoEL [82], MIME [89], EmpDG [74], KEMP [75], CEM [113], RecEC_{soft} [44], and GEE_{MIME} [64]. The above models

can be divided into two categories. The first four models consider only the affection aspect of empathy, i.e., the user’s emotions and have been introduced in Section 3.3.1. The last four models also take into account cognitive empathy:

- **KEMP**. Utilizes external commonsense knowledge and an emotional lexicon to better comprehend and generate empathetic responses [75].
- **CEM**. Produces the causes and effects of the user’s most recent experiences and incorporates them into context encoding to enhance understanding of the user’s situations and emotions [113].
- **RecEC_{soft}**. Focuses on emotional causes extracted from the dialogue context at the word level using a soft gated attention mechanism within the decoder [44].
- **GEE_{MIME}**. Applies a rational speech act framework to refine MIME’s generated response, emphasizing words that trigger emotions in the dialogue context [64].

All the above models, as well as ours, are built upon a transformer backbone for a fair comparison. For a fair comparison, the parameter settings and training processes of all comparison models are the same as suggested in their original papers and implementations.

Implementation Details *Pytorch* is used to implement the proposed models. The token embedding is initialized with pre-trained Glove vectors of dimension 300 [103] and shared by the transformer encoder, graph-based encoder, and the projection layer of the decoder. The hidden size d is set to 300. The number of nodes $|\mathcal{V}|$ is 800, and the number of selected relationships k is 512 (0.16%). Both the encoder layer number and the decoder layer number are 2. The batch size is

	Model	PPL↓	BLEU-3↑	BLEU-4↑	P _{BERT} ↑	R _{BERT} ↑	F _{BERT} ↑
Aff	MoEL	36.87±0.97	4.53±0.53	2.80±0.32	.499±.008	.467±.007	.480±.006
	MIME	37.88±0.49	4.48±0.15	2.71±0.09	.490±.004	.466±.002	.475±.002
	EmpDG	55.64±3.78	3.64±0.38	1.99±0.22	.475±.007	.458±.008	.465±.004
Aff+	KEMP	36.59±0.45	4.13±0.29	2.43±0.15	.484±.005	.460±.004	.470±.005
	CEM	36.70±0.44	3.55±0.42	2.24±0.24	.498±.001	.461±.006	.477±.004
	RecEC _{soft}	149.3±15.9	3.02±0.15	1.62±0.12	.491±.004	.461±.002	.473±.002
Cog	GEE _{MIME}	-	2.76±0.18	1.50±0.14	.472±.002	.443±.002	.456±.001
	GREC	32.66* ±0.32	4.96* ±0.14	3.16* ±0.08	.532* ±.003	.506* ±.003	.512* ±.004
	CARE	32.84* ±0.23	4.88* ±0.13	2.95* ±0.06	.501 ±.004	.475* ±.002	.486* ±.003

Table 4.1: CARE: Automatic evaluation results.

set as 16. When training the model, we use Adam optimizer [65] and vary the learning rate following [129].

4.3.2 Evaluation

Automatic Evaluation We employ PPL, BLEU, and BERTScore to automatically assess the performance of the proposed method. The evaluation results on the EMPATHETICDIALOGUE test set are presented in Table 4.1. Here, *Aff* and *Cog* indicate models that incorporate affective and cognitive aspects for empathy, respectively. For each method, experiments are repeated multiple times using different random seeds, and the reported values correspond to the mean along with the standard deviation. Results marked with * denote statistical significance at $p < 0.05$. For each metric, the highest score is highlighted in bold.

According to the results, CARE outperforms all models except GREC in terms of all metrics. The lowest perplexity score suggests that our proposed architecture is more confident in its generated responses than other models. The table does not present the perplexity score of GEE_{MIME}. This is because its generated token probability distribution depends on the mediated results of MIME and its emotion cause detector, and therefore, PPL is less relevant to its core structure, i.e., rational speech act framework. The highest BLEU and BERTScore

scores indicate that our approach can generate more human-like responses by incorporating causality reasoning. Especially, all the above advantages are significant and stable, evident in high degrees of statistical significance and small standard deviations, respectively. Comparing the GREC model to others, we found that external knowledge significantly improves response quality. This effect is especially pronounced when measured against the golden response.

Human Ratings Although automatic evaluations provide insights into model performance, they are limited in capturing aspects such as empathetic understanding and contextual appropriateness. Following prior work, we randomly select 128 conversations along with their generated responses from different models for human assessment. Three PhD students with expertise in mental health support conducted the evaluation. For each generated response, they rated the following dimensions: (1) *Empathy*, indicating the extent to which the response reflects understanding of the user’s feelings and experiences; (2) *Relevance*, assessing whether the response is on-topic and suitable in the context of the preceding conversation; (3) *Fluency*, evaluating the linguistic quality and coherence of the response. All ratings were made on a 5-point Likert scale, where 5 represents the highest quality. The final score for each metric was obtained by averaging the three annotators’ ratings.

Table 4.2 presents the results of human evaluations, with the highest scores highlighted in bold. To quantify agreement among annotators, we compute Fleiss’s kappa, obtaining a value of 0.41, which indicates a moderate level of inter-rater consistency. From these results, two key observations emerge. First, CARE achieves the top scores in Empathy and maintains relatively high Relevance and Fluency compared to most baseline models. This suggests that our interdependent and simultaneous causal reasoning indeed enhances both em-

Model		Empathy \uparrow	Relevance \uparrow	Fluency \uparrow
Affection	MoEL	2.73	2.63	4.82
	MIME	2.30	2.24	4.88
	EmpDG	2.31	2.27	4.52
Affection +Cognition	KEMP	2.26	2.18	4.81
	CEM	2.77	2.70	4.93
	RecEC _{soft}	2.16	2.21	4.74
	GEE _{MIME}	1.75	1.75	4.78
	GREC	2.79	2.87	4.84
	CARE	2.83	2.79	4.86

Table 4.2: CARE: Human rating results.

pathetic expression and content relevance, as intended. By leveraging reasoned causalities, CARE better comprehends user emotions and experiences. Moreover, this reasoning allows the model to infer plausible user feelings and experiences that are not explicitly stated in the prior conversation, enabling it to generate responses with stronger empathy, as illustrated in the case study. Second, models that incorporate both affective and cognitive factors (shown in the bottom half of the table) do not consistently outperform those focusing solely on affective aspects (top half of the table). This pattern is also reflected in the automatic evaluation results in Table 4.4. Although causal reasoning can intuitively aid the understanding of user feelings and experiences, unrefined reasoning may result in a narrow perspective and diminished empathetic quality. Comparing GREC and CARE, we observe that external knowledge can improve response relevance, yet grounding empathetic understanding in the dialogue context contributes more significantly to the expression of empathy.

Human A/B Test To assess the overall performance of various models, we conduct a human A/B evaluation. For a pair of models, A and B, in our study, *CARE vs. Multi-TRS, MoEL, MIME, EmpDG*, annotators are asked to judge which model generates better responses, imagining themselves as the users. If the responses from both models are equally good or bad, annotators can choose

CARE vs.	CARE wins	CARE loses	Tie
Multi-TRS	51.7%	20.0%	28.3%
MoEL	50.0%	21.7%	28.3%
MIME	41.7%	28.3%	30.0%
EmpDG	68.3%	10.0%	21.7%

Table 4.3: CARE: Human A/B test results.

Model Variant	PPL↓	BLEU-3↑	BLEU-4↑	P _{BERT} ↑	R _{BERT} ↑	F _{BERT} ↑
w/o reasoning	33.34	4.74	2.83	.493	.473	.481
w/o condition	33.23	4.74	2.83	.501	.473	.485
Full model	32.84	4.88	2.95	.501	.475	.486

Table 4.4: CARE: Automatic evaluation results of the ablation study.

Model Variant	Empathy↑	Relevance↑	Fluency↑
w/o reasoning	2.38	2.23	4.86
w/o condition	2.60	2.47	4.87
Full model	2.83	2.79	4.86

Table 4.5: CARE: Human rating results of the ablation study.

Tie. As shown in Table 4.3, responses produced by CARE are generally favored by human judges compared to those from Multi-TRS, MoEL, MIME, and EmpDG. These results suggest that CARE can generate responses that are not only more empathetic but also more contextually relevant than models that focus solely on the user’s emotions.

4.3.3 Further Analyses

Ablation Study For deeper analyses of CARE, we investigate its inner structures and functions. In specific, we propose two variant models to verify the contribution of reasoning and the reasoning condition in CARE:

- (1) **w/o reasoning:** We remove the CVGAE structure and directly incorporate the prior causal graph into response generation.
- (2) **w/o condition:** We replace CVGAE with VGAE to eliminate the effect of

the reasoning condition.

Results are shown in Table 4.4 and Table 4.5, respectively. From Table 4.4, both variants achieve relatively high automatic evaluation metric scores. Moreover, the variant models surpass previous comparison models in Table 4.1. It indicates that causalities can help models respond more like humans, given that both variants consider additional causalities between the user's experiences. However, both variants' performances in terms of human evaluation are relatively low. Accordingly, we can draw the following three conclusions:

- (1) Not all information in the golden response contributes to empathy. Although two variants have high automatic evaluation scores, they fail to achieve equally high human ratings. Such a phenomenon can also be clearly observed when comparing the performance of EmpDG and KMEP.
- (2) Generated responses considering causalities not mentioned in dialogue history are more empathetic and relevant, which is supported theoretically by [106]. It is reflected by higher **Empathy** and **Relevance** of **w/o condition** than **w/o reasoning**, given that **w/o condition** reasons plausible causalities not mentioned in dialogue history compared with **w/o reasoning**.
- (3) Emotional and contextual information guides the model to reason causalities contributing to empathetic expression, given that the full model has higher Empathy and Relevance than **w/o condition**.

Analysis of #SelectedRelationships k Fig. 4.3 demonstrates the model performance when we increase the number of selected relationships k . The solid and dotted lines represent BLEU-4 and a two-period moving average, respectively.

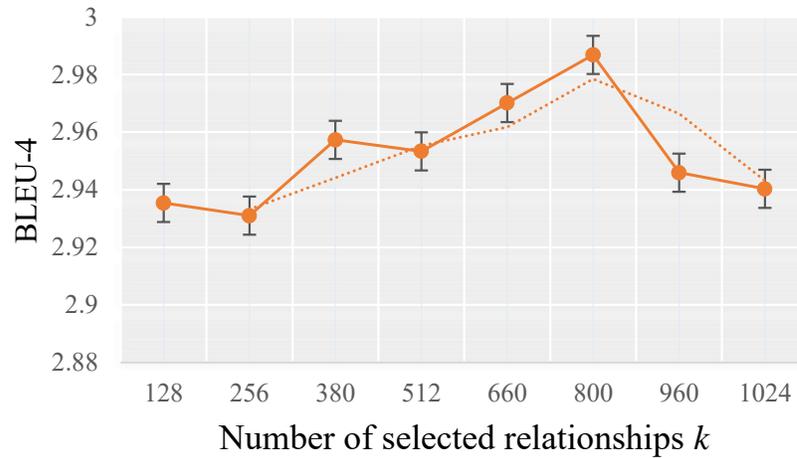


Figure 4.3: CARE: Model performs with the number of selected relationships k increasing.

For each k , we repeat five runs and compute the average BLEU-4. The performance of CARE with regard to BLEU-4 first rises and then drops as we increase the number of relationships infused in response generation k . It indicates that sufficient causalities benefit empathetic expression, but excessive ones could involve noise and hurt empathy.

Case Study Table 4.6 presents a case along with responses generated by the models and comparison models. The content showing comprehension of feelings and experiences is highlighted in italic. From the table, CARE can respond more empathically to the user than other models. Notably, CARE is able to show deep and considerate comprehension of the user’s feelings and experiences in the response. For instance, it understands that the “apprehensive” emotion comes from the lack of *confidence* and the user has already proposed a quite effective solution (*great idea*).

Speaker Emotion: apprehensive	
Speaker:	I had some hesitations when I was at the gym today, thought I wouldn't be able to handle the weights.
Listener:	How'd it turn out? Congrats on deciding to lead a healthier life, by the way.
Speaker:	I decided to focus on the music that was blasting, just tried switching up what I play in the gym, it sounds like vikings are bringing me into battle
MoEL	That is great. I am sure you will do great!
MIME	I am sure you will be fine.
EmpDG	That is a great time to have! You should always feel the same way.
KEMP	I am glad you are going through that.
CEM	I am sure it will be fun!
RecEC _{soft}	That is a lot of money. How long did you do?
GEE	I hate that. I hope you are okay.
CARE	That is a <i>good idea</i> . I hope you have a lot of <i>confidence</i> .

Table 4.6: CARE: Case study.

4.4 Conclusion

This chapter also focuses on empathetic understanding of users via emotional causalities. In particular, a method is proposed to reason about all plausible emotional causalities in conversations interdependently and simultaneously, aiming for a deeper understanding of the user's feelings and experiences in empathetic dialogue. This problem is formulated as a conditional graph generation task. To address it, we design CVGAE, a model that leverages latent variables for conditional structure prediction, incorporating both the dialogue context and predicted future conversation content to facilitate causality reasoning. The inferred causalities are then integrated into the response generation process through a multi-source attention mechanism in the decoder. The complete framework is referred to as CARE (CAusality Reasoning for Empathetic conversation). Experimental results on the EmpatheticDialogues benchmark demonstrate that CARE outperforms previous methods in both automatic and human evaluations. Compared with GREC, introduced in the previous chapter, CARE conducts emotional causality reasoning through its internal model structure, thereby eliminating the reliance on external knowledge bases.

Part II

Reliable Responding

The empathetic understanding of emotionally intelligent conversational agents has been explored. Building on this foundation, this part further investigates the challenge of generating reliable and supportive responses in emotional support conversations. Emotional support conversation systems are designed to generate responses that can buffer the emotional distress experienced by users (help-seekers) and help users to work through the challenges they are confronting [85]. Prior studies have enhanced response generation by incorporating dialogue strategies [164, 23] and/or leveraging external knowledge sources [34, 128], aiming to improve the overall supportiveness and empathy of system outputs. These systems are typically designed to act as supporters, providing emotional buffering for users (help-seekers) and assisting them in navigating the challenges they face [85]. Despite significant progress, ensuring the reliability of responses from emotionally supportive systems remains a critical challenge. As these systems interact with a wide range of users, an important open problem is how to satisfy diverse individual needs while avoiding harmful or inappropriate responses. Addressing this issue is essential for building effective and reliable emotional support systems. Thus, two methods are proposed in this part to improve user satisfactory and mitigate unhelpful responses, respectively.

Chapter 5

User Satisfactory Improvement via a Bayesian Approach

5.1 Introduction

Maximizing user satisfaction across diverse individuals is one of the important aspects of reliable responding. It is a common method to align models with human preferences using either contrastive learning or reinforcement learning. In this process, human preference learning is significant. Due to the high subjectivity of personal standards and human values, it can be difficult to reach a consensus on preferences among individuals in scenarios of emotional support, significantly increasing the learning difficulty. As depicted in Fig. 5.1, people can have different feelings towards the same response in the emotional support conversation because of their own experiences and values. A reliable, emotionally intelligent conversational agent is expected to consider the benefits of universal groups, including minorities, and generate less controversial and more helpful content, like supporter B instead of A. Existing methods for tackling this challenge often rely on aggregation techniques like majority voting or averaging to reconcile human preferences [58, 38, 13]. However, this approach has two key limitations. First, aggregated preferences may cater to specific sub-

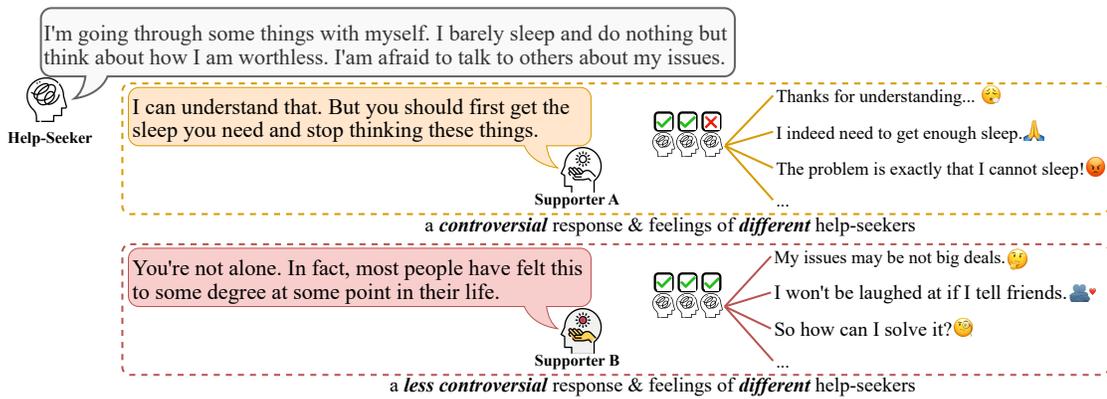


Figure 5.1: Alignment with d-PM: Different levels of user satisfaction.

groups, risking the generation of controversial content as seen with Supporter A in Figure Fig. 5.1. Second, certain tasks, such as emotion analysis [68] and ethical evaluation [51], are inherently subjective and inconsistent [9], making dismissal of this disagreement problematic. Therefore, to generate widely acceptable and less controversial outputs, it is crucial for a model to account for the disagreements inherent in human preferences [71], rather than simply dismissing or averaging them out. To solve this problem, we introduce a novel Bayesian-based approach termed Preference Modeling with Disagreement (**d-PM**). This method is designed to approximate a “universal preference” that comprises the preferences of “all individuals,” given the preferences of several individuals. Although a soft label, derived from several individuals, can intuitively account for disagreement, outliers, or extreme labels can disproportionately influence the overall perception. Therefore, we employ Bayesian inference to refine these preferences. The observed preferences from a selected group of individuals serve as prior knowledge. Our proposed **d-PM** (distributional preference modeling) framework leverages the distribution of all possible universal preferences, or the likelihood probability, to adjust and smooth this initial data. This process allows us to derive a universal preference, a posterior distribution

that is more robust and less biased. Once the universal preference is obtained, we calculate a preference score for a given text by evaluating its likelihood under the expected preference types. This score is then used to align a language model. Based on our **d-PM** model, we further optimize language models to generate widely acceptable and less controversial texts. We propose using a contrastive learning strategy to calibrate the generation model toward producing responses with high preference scores. This approach is more advantageous than existing RL strategies, which are often computationally costly in terms of convergence [47] and online decoding [165].

The main contributions are threefold:

- (1) To the best of our knowledge, this work is the first to align emotional support conversational agents with human preferences while considering inherent disagreement among different individuals.
- (2) In order to model human preferences with their disagreement, a Bayesian approach is proposed, Preference Modeling with Disagreement (d-PM). Additionally, we use its preference scores to calibrate emotional support conversational agents via contrastive learning for generations that can be widely accepted and less controversial.
- (3) We conduct experiments on an emotional support benchmark, i.e., emotional support conversation, to validate the performance of the proposed method. Experimental results demonstrate the effectiveness and versatility of the proposed method.

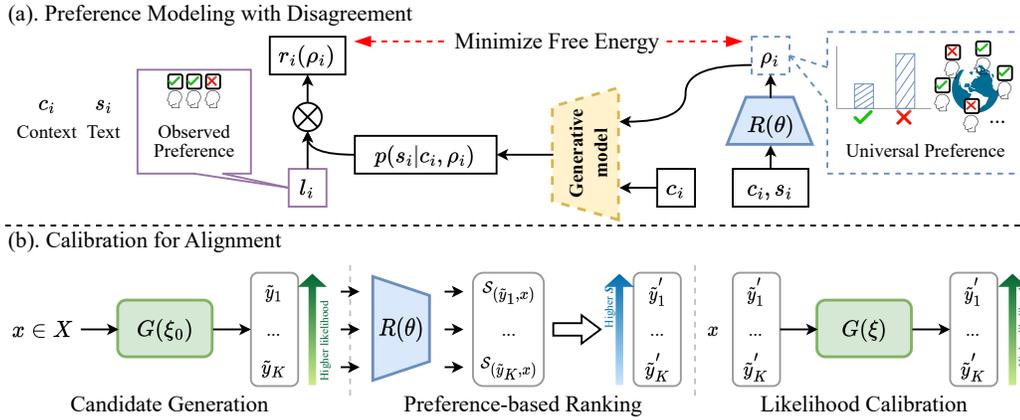


Figure 5.2: Alignment with d-PM: Preference modeling with disagreement and alignment accordingly.

5.2 Method

5.2.1 Problem Formulation

Fig. 5.2 shows the overview of the proposed method. We first use a Bayesian approach, i.e., d-PM, to model human preferences with disagreement (Section 5.2.2). Then, we calibrate an emotional support conversational agent by contrastive learning with preference scores of d-PM to align this model with human preferences (Section 5.2.3). Generally, these two processes can be formulated as follows.

Preference Modeling with Disagreement Human preferences can be inferred from a human-annotated dataset, denoted as \mathcal{D} . Each instance in the dataset is represented as a triplet (c, s, l) . Here, c is a context; s is a text; and l is a label indicating the annotators' preferences. The inherent disagreement in human preferences can be encapsulated within the label in two distinct ways. In the first approach, the label can be a soft label derived from multiple annotations, all attributed to the same sentence [68]. These annotations are sourced

from multiple human annotators to preserve disagreement among individuals. The second approach is the direct collection of a global consensus. In this context, the label signifies the proportion of people who find a particular sentence acceptable, an estimate provided by a single human annotator [172, 41].

Aimed at capturing human preference with disagreement within the dataset, we assume there is a distribution ρ over two classes, $\{acceptable, unacceptable\}$, comprising preferences of all humans, and therefore l is the sampling result from ρ . We employ a preference model $\mathcal{R}(\theta)$ to infer ρ given c and s as inputs and the probabilistic format of l as the prior distribution. Since we focus on whether s is widely acceptable, the likelihood of the class *acceptable* is defined as the preference score:

$$\mathcal{S}_{(s,c)} = \mathcal{R}(s, c; \theta)_{acceptable}. \quad (5.1)$$

Calibration for Alignment To align emotional support models with human preferences, the existing generation model $\mathcal{G}(\xi_0)$ is calibrated using preference scores. Here, $\mathcal{G}(\xi_0)$ denotes a model that has been fine-tuned on a dataset (X, Y) , where X and Y represent the input set and the corresponding output set, respectively, and ξ_0 are the optimized parameters. Significantly, if the dataset for preference modeling is identical to the (X, Y) dataset, then $(x, y) \in (X, Y)$ corresponds to $(c, s) \in \mathcal{D}$. Otherwise, the two datasets should belong to a similar domain to ensure compatibility. For each $x \in X$, K candidate sequences $\{\tilde{y}_k\}_{k=1}^{k=K}$ are decoded using $\mathcal{G}(\tilde{y}_k|x; \xi_0)$. The model $\mathcal{G}(\xi_0)$ is then further optimized to align the likelihoods of these candidate responses with their associated preference scores $\{\mathcal{S}_{(\tilde{y}_k, x)}\}_{k=1}^{k=K}$, thus encouraging generation behaviors that reflect human-aligned preferences.

5.2.2 Preference Modeling with Disagreement

A distribution ρ is established to represent the universal preference for the text s given its context c . Therefore, the observed annotations l are considered as samples from ρ , and can form a prior distribution $p_i(\rho)$. Drawing inspiration from [111], a Bayesian approach is adopted to approximate ρ using this prior.

Specifically, a connection between ρ and l is established through the optimization process of a generative model. This model is designed for generating text s conditioned on c_i and ρ : $p(s|c_i, \rho)$. The log-likelihood of the text can be formulated as $\sum_i \log p(s_i|c_i) = \sum_i \log (\sum_\rho p(s_i|c_i, \rho) p_i(\rho))$, where $p_i(\rho)$ is the prior preference distribution. Its optimization can be achieved by introducing a variational posterior distribution $q(\rho|s_i, c_i)$ for the i -th datapoint, and minimizing the free energy (negated evidence lower bound) formulated as:

$$-\sum_i \log p(s_i|c_i) + \sum_i \sum_\rho q(\rho|s_i, c_i) \log \frac{p(s_i|c_i, \rho) p_i(\rho)}{q(\rho|s_i, c_i)}. \quad (5.2)$$

Minimization of the free energy involves estimations of both the forward distribution of text s_i : $p(s_i|c_i, \rho)$, and the posteriors $q(\rho|s_i, c_i)$, which can be computed by our preference model:

$$q(\rho|s_i, c_i) = \mathcal{R}(s_i, c_i|\theta). \quad (5.3)$$

As for $p(s_i|c_i, \rho)$, it is defined only on the i -th datapoint and is computed by minimizing Eq. (5.2) for fixed $q(\rho|s_i, c_i)$, s.t., $\sum_i p(s_i|c_i, \rho) = 1$ for all ρ . Thus, the optimum is achieved by:

$$p(s_i|c_i, \rho) = a_{i,\rho} = \frac{q(\rho|s_i, c_i)}{\sum_j q(\rho|s_j, c_j)}. \quad (5.4)$$

From Eq. (5.4), the generative model can be regarded as a matrix of variables $a_{i,\rho}$ describing conditional probabilities of different responses s_i given different latent distributions ρ with known c_i .

In a variational way, Eq. (5.2) can be rewritten as $-\log p(s_i|c_i) + \sum_i \text{KL}(q(\rho|s_i, c_i) \| r_i(\rho))$. Here, $r_i(\rho) \propto p(s_i|c_i, \rho)p_i(\rho)$ is the posterior model of the generative model, and it can be reformulated with reduction of $p(s_i|c_i, \rho)$ to the matrix in Eq. (5.4):

$$r_i(\rho) = \alpha_i \cdot p_i(\rho)p(s_i|c_i, \rho) = \alpha_i \frac{p_i(\rho)q_i(\rho|s_i, c_i)}{\sum_j q_j(\rho|s_j, c_j)}, \quad (5.5)$$

where α_i is a scalar enabling $\sum_\rho r_i(\rho) = 1$. Accordingly, minimizing the free energy reduces to minimizing the KL divergence between $q(\rho|s_i, c_i)$ and $r_i(\rho)$. The minimization of the free energy in Eq. (5.2) can be derived as:

$$\sum_i \text{KL}(q(\rho|s_i, c_i) \| r_i(\rho)) = \min_\theta \sum_i \text{KL} \left(\mathcal{R}(s_i, c_i; \theta) \left\| \alpha_i \cdot p_i(\rho) \frac{\mathcal{R}(s_i, c_i; \theta)}{\sum_j \mathcal{R}(s_j, c_j; \theta)} \right. \right). \quad (5.6)$$

By optimizing the above objective, the parameters θ of the preference model can be effectively learned.

5.2.3 Calibration for Alignment

$\mathcal{G}(\xi_0)$ is able to generate texts with both high and low preference scores. As shown in Fig. 5.3, it demonstrates evaluations over 1,000 data instances across three emotional support conversation models. However, we expect to calibrate the model such that the generation probability aligns with these preference scores. Specifically, diverse beam search [131] is used to generate multiple candidates and then use d-PM to evaluate these candidates. For the sake of more likely generating a high preference score text, we propose a model-agnostic module to leverage contrastive learning to calibrate generation likelihood, aligning with d-PM. Taking inspiration from recent calibration work [124, 86, 161], this process is implemented through the following three steps:

Step 1: Candidate Generation We generate candidates from the text generator $\mathcal{G}(\xi_0)$, which has been fine-tuned on corresponding dataset (X, Y) and its parameters are ξ_0 , on its own training dataset. Given an input sequence $x \in X$, we first use $\mathcal{G}(\xi_0)$ to generate K candidates $\{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_K\}$ using diverse beam search. As a result, these candidates will get similar possibilities yet different preference scores according to the above preliminary study.

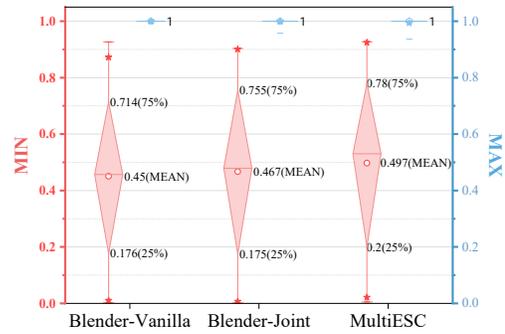


Figure 5.3: Alignment with d-PM: The maximum and minimum preference scores of 10 candidates generated via diverse beam search.

Step 2: Preference-based Ranking We use our proposed d-PM $\mathcal{R}(\theta)$ to measure the preference score $\mathcal{S}_{(\tilde{y}_k, x)}$ of each candidate \tilde{y}_k . Then we rank these candidates according to the above preference score and obtain a list of ranked candidates: $\tilde{y}'_1, \tilde{y}'_2, \dots, \tilde{y}'_K$, where $\mathcal{S}_{(\tilde{y}'_i, x)} > \mathcal{S}_{(\tilde{y}'_j, x)}$ for $\forall i < j$.

Step 3: Likelihood Calibration As mentioned before, we leverage contrastive learning to assign higher likelihoods to the candidates with higher preference scores. The following pairwise margin loss is used to adjust the generator $\mathcal{G}(\xi)$.

$$\mathcal{L}^r = \sum_i \sum_{j>i} \max(0, \mathbf{P}(\tilde{y}'_j; \xi) - \mathbf{P}(\tilde{y}'_i; \xi) + \lambda_{ij}), \quad (5.7)$$

where λ_{ij} is the default margin λ multiplied by the difference in rank between the samples, i.e., $\lambda_{ij} = \lambda * (j - i)$. $\mathbf{P}(\tilde{y}'_i; \xi)$ is the length-normalized log-probability of the candidate:

$$\mathbf{P}(\tilde{y}'_i; \xi) = \frac{\sum_{t=1}^{|\tilde{y}'_i|} \log \mathcal{G}(\tilde{y}'_t | x, \tilde{y}'_{<t}; \xi)}{|\tilde{y}'_i|^\alpha}, \quad (5.8)$$

where α is the length penalty hyperparameter. To avoid forgetting the token-level likelihood information of the ground-truth text, we also use an additional token-level negative log-likelihood. The final calibration loss is as follows:

$$\mathcal{L}^c = -\lambda \frac{1}{|y|} \sum_{t=1}^{|y|} \log \mathcal{G}(y_t | x, y_{<t}; \xi) + \mathcal{L}^r. \quad (5.9)$$

We minimize \mathcal{L}^c to optimize the generator’s parameters ξ . This process is supervised by our d-PM model and aligns the generation model with human preference.

5.3 Experiments

5.3.1 Experimental Settings

ESConv dataset (\mathcal{D}) The models are tested on the task of emotional support conversation. The benchmark ESConv [85] is adopted to train and test the models’ performance. This dataset contains approximately *1k conversations with 31k utterances*, developing each conversation between a help-seeker and a supporter. All base models follow the original division of ESConv for training, validation, and testing, using an 8:1:1 ratio. Notably, each model adopts different data pre-processing methods. We adhere to each base model’s specific data division and pre-processing.

d-PM Training The d-PM models are built by fine-tuning the BERT (base, uncased) model with prefix-tuning. BERT consists of $n_{\text{layer}}^{\text{BERT}}$ layers with hidden size d^{BERT} . A sequence of prefix tokens, denoted by L_{idx} , of length $|L_{\text{idx}}|$ is prepended to the input. The model input is formed by concatenating the prefix, the conversation context, and the text: $z = [L_{\text{idx}}; c; s]$, with z_t representing

the input at time step t . Correspondingly, the activations are $h = [h^{\text{prefix}}, h^c, h^s]$, where h_t stacks all layer outputs at step t . Here, h^c and h^s are computed using the original BERT parameters, while h^{prefix} consists of learnable parameters. A trainable matrix A_{θ_1} , parameterized by θ_1 , is initialized with size $|\text{L}_{\text{idx}}| \times d^{\text{dim}}$, where $d^{\text{dim}} = n_{\text{layer}}^{\text{BERT}} \times d^{\text{BERT}}$. Thus, the activation h_t at each time step t can be expressed as:

$$h_t = \begin{cases} A_{\theta_1}[t, :] & \text{if } t < |\text{L}_{\text{idx}}|, \\ \text{BERT}(z_t, h_{<t}) & \text{otherwise.} \end{cases} \quad (5.10)$$

We employ the final hidden states of the text s (i.e., the outputs from BERT’s last layer) to compute the probabilities for the classes *acceptable* and *unacceptable*. This is achieved through a multilayer perceptron (MLP) parameterized by θ_2 .

The overall parameters of the d-PM, $\theta = (\theta_1, \theta_2)$, are trained using the loss defined in Eq. (5.6). The preference models used in Alignedmajor and Alignedsoft adopt the same network architecture, but differ in their respective loss functions, as specified in Eq. (5.11) and Eq. (5.12). Our d-PM implementation is based on the publicly released code of the UnifiedSKG framework [155]. During training, BERT’s parameters were kept frozen, and only the prefix was updated. We set the prefix length to 10, used a batch size of 160, and a learning rate of 5×10^{-4} .

The human preference labels are obtained from the Motivational Interviewing Dataset (MI-Dataset) [143], which contains approximately 17k supporter responses to help-seeker utterances. Each response was annotated by 2 \sim 4 experts according to the MI codes [92]. These annotations can be mapped into two classes, $\{\textit{acceptable}, \textit{unacceptable}\}$, and disagreements among annotators are estimated using our d-PM approach.

To build the d-PM, a BERT model is fine-tuned on the MI-Dataset with prefix-

The Classes	Coarse-Grained Codes	Fine-Grained Codes	Examples
<i>acceptable</i>	MI Adherent	1. Advise with Permission	If you agree with it, I could try to brainstorm some ideas that may help.
		2. Affirm	You should be proud of yourself for your past efforts.
		3. Emphasize Autonomy	It is really up to you to decide.
		4. Support	I know it's really hard to stop drinking.
	Relational	5. Closed Question	Do you think this is an advantage?
		6. Open Question	What is your take on that?
		7. Simple Reflection	It sounds like you're feeling worried.
		8. Complex Reflection	Speaker: Mostly, I would change for future generations. Listener: It sounds like you have a strong feeling of responsibility
		9. Give Information	Logging your cravings is important as cravings often lead to relapses.
		10. Self-Disclose	I used to be similar where I get obsessed about how people look.
		11. Other	Good morning, Hi there.
<i>unacceptable</i>	MI Non-Adherent	12. Advise without Permission	You should simply scribble a note that reminds you to take a break.
		13. Confront	Yes, you are an alcoholic. You might not think so, but you are.
		14. Direct	Don't do that!
		15. Warn	Be careful, DO NOT stop taking meds without discussing with your doctor.

Table 5.1: Alignment with d-PM: The MI codes in MI-Dataset.

tuning [155, 76]. This dataset originally contains 15 fine-grained response classes, which are further grouped into three coarse-grained categories based on MI codes [92]: (1) *MI Adherent*: responses providing empathetic, compassionate support, helping the help-seeker feel understood and respected; (2) *Relational*: responses aimed at fostering a positive relationship between the help-seeker and supporter, promoting more constructive interactions; (3) *MI Non-Adherent*: responses such as arguing, confronting, or giving unsolicited advice, which may generate resistance and impede problem-solving.

For training the d-PM, these coarse-grained categories were converted into the binary classes *acceptable* and *unacceptable* (see Table 5.1). The dataset was randomly split into training and validation sets with a 9 : 1 ratio. Approximately 34.93% of the instances exhibit annotation disagreement among the experts.

Emotional Support Model Training The following models are adopted as the base models $\mathcal{G}(\xi_0)$ of the aligned emotional support model:

- Vanilla refers to a standard BlenderBot [112] trained on the ESConv dataset. The small version of BlenderBot is used in experiments, consistent with prior studies.
- Joint is built on the BlenderBot backbone and is specifically trained to generate responses prefixed with an expected communication skill [85]. This model primarily emphasizes skill efficiency.
- MultiESC [23] is an emotional support conversation model that also focuses on communication skill efficiency. It predicts communication strategies for several upcoming turns based on the user’s emotional state.

We apply our proposed method to align each base model, thus treating the well-trained base model as the generator $\mathcal{G}(\xi_0)$. To evaluate the effectiveness of d-PM, three alternative preference models are implemented and integrated into the framework for comparative analysis:

- (1) A preference model (major) trained to predict the majority voting result of annotations from different annotators, denoted as l_m , and optimized by cross-entropy loss, formulated as:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(c,s,l_m) \sim \mathcal{D}} [p_l(l_m) \log(\mathcal{R}^{l_m}(c, s; \theta))]. \quad (5.11)$$

where $p_l(l_m)$ denotes the one-hot vector of l_m .

- (2) A preference model (soft) trained to approximate the direct probabilistic label of annotations, i.e., the soft label l . The model is optimized by:

$$\mathcal{L}(\theta) = \mathbb{E}_{(c,s,l) \sim \mathcal{D}} \|\mathcal{R}(c,s;\theta) - l\|^2. \quad (5.12)$$

- (3) A preference model (w/oA) that does not aggregate annotations and takes each annotation as independent. This model is optimized by cross-entropy loss, similar to Eq. (5.11).

We followed the implementation procedures of the base models as described in [85, 23] and utilized their publicly released code. Each base model was executed once using the seed specified in its code. To ensure a fair comparison, the aligned models were configured with hyperparameters matching those of their respective base models, with the exception of an increased learning rate to improve training efficiency. Concretely, the learning rate was set to 1×10^{-3} for Blender-Vanilla and Blender-Joint, and 3×10^{-5} for all other models. Due to GPU memory limitations, the batch size was reduced from 32 to 12 when training the aligned MultiESC model. For robustness, each aligned model was trained five times using different seeds (0, 1, 13, 42, 1024), and the reported results are the averages across these runs.

When training the aligned models, we aim to retain the same hyperparameters used in the training of the base models. We set the candidate number K to 10. We train each aligned model five times with five different seeds. Subsequently, we test each of the five trained models on the test dataset and compute the average results.

Model		B-1	B-2	B-3	B-4	R-L	METEOR	CIDEr	Extreme
Blender -Vanilla	Base	17.85	7.08	3.60	2.11	17.06	7.46	15.44	51.02
	Aligned _{major}	19.07	7.71	3.94	2.28	17.09	7.71	15.97	50.61
	Aligned _{soft}	17.88	7.21	3.68	2.12	16.52	7.31	15.50	50.73
	Aligned _{w/oA}	19.70	7.56	3.64	2.05	16.90	7.72	15.62	50.48
	Aligned _{d-PM}	20.75	8.32	4.17	2.39	17.41	8.21	16.57	50.38
Blender -Joint	Base	18.70	7.30	3.61	2.03	17.66	7.56	16.91	50.95
	Aligned _{major}	20.37	8.61	4.47	2.65	19.23	8.32	21.86	51.57
	Aligned _{soft}	19.36	7.87	3.85	2.09	17.55	7.65	15.90	50.84
	Aligned _{w/oA}	21.05	8.14	3.89	2.07	17.65	8.11	15.29	50.68
	Aligned _{d-PM}	21.05	8.97	4.74	2.78	19.39	8.48	20.34	51.81
MultiESC	Base	20.36	8.80	4.92	3.14	21.00	8.58	30.69	52.74
	Aligned _{major}	19.10	8.27	4.61	2.88	20.72	8.24	30.15	52.57
	Aligned _{soft}	19.30	8.33	4.62	2.88	20.83	8.35	30.75	52.54
	Aligned _{w/oA}	21.58	8.80	4.74	2.96	20.47	8.78	28.58	51.65
	Aligned _{d-PM}	21.59	9.56	5.33	3.36	21.50	9.03	32.65	53.15

Table 5.2: Alignment with d-PM: Automatic evaluation results.

5.3.2 Evaluation

Automatic Evaluation We adopt the following metrics commonly used in previous work [23, 85] for the automatic evaluation of our proposed method: BLEU [100] (B-1/2/3/4), ROUGE (R-L) [80], METEOR [6], CIDEr [130], and BOW Embedding-based matching score [84] (Extreme).

The results are summarized in Table 5.2. Across all automatic metrics, the Aligned_{d-PM} model consistently outperforms its corresponding base model, with statistical significance at $p < 0.01$. This improvement holds regardless of the underlying base model, demonstrating the general effectiveness of our proposed approach. Aligned_{major} and Aligned_{soft} can enhance performance for Blender-Vanilla and Blender-Joint, but they fail to provide gains for MultiESC. This highlights the limitations of using majority-vote or soft labels to reconcile discrepancies in human preferences. Aligned_{w/oA} shows mixed results: it surpasses the base model on certain metrics but falls short on others. Its CIDEr score, which evaluates similarity using TFIDF-weighted n-grams, is notably lower, indicat-

ing that it is less capable of producing responses containing critical information from the reference. This limitation arises because the preference scores from w/oA are narrowly distributed, limiting its ability to correctly rank generated outputs. Taken together, these findings emphasize the advantage of our proposed preference model, d-PM, in capturing human-aligned preferences more effectively.

Human Evaluation We conducted human evaluations on model generations using MultiESC, as it consistently outperforms Blender-Vanilla and Blender-Joint across most automatic metrics. Specifically, we randomly selected 100 model-generated responses for human assessment. Six annotators with diverse educational backgrounds, genders, geographic regions, and occupations participated in the evaluation. Each response was rated by three different annotators, and the final score was computed as the average of their ratings. For each input, 3–4 responses generated by different models were presented together, along with the corresponding context or prompt-reply pair. Annotators were instructed to compare these responses relative to each other. To mitigate potential bias, the order of the sentences was randomized, preventing any inference of the generating model based on position.

Annotators were asked to put themselves in the position of the help-seeker and evaluate each response across five dimensions: (1) *Identification*: rated on a scale from 1 to 5, assessing how well the response helps explore the situation and pinpoint problems; (2) *Comforting*: 1–5 scale evaluating the response’s ability to provide emotional comfort; (3) *Suggestion*: 1–5 scale measuring how effectively the response offers practical advice or solutions; (4) *Overall*: 1–5 scale for the general quality of the response as emotional support; (5) *Global Consensus*: number of annotators agreeing that the response is helpful, with 1–5 cor-

Model	Identification	Comforting	Suggestion	Overall	Global Consensus
Base	3.017	2.562	2.918	2.598	2.693
Aligned _{major}	3.032	2.572	2.880	2.598	2.763
Aligned _{soft}	3.007	2.557	2.905	2.568	2.747
Aligned _{d-PM}	3.052	2.587	2.952	2.637	2.783

Table 5.3: Alignment with d-PM: Human evaluation results on ESConv.

responding to nobody ($< 1\%$), rare ($5\% \sim 25\%$), controversial ($\sim 50\%$), most ($75\% \sim 90\%$), and all ($> 99\%$). Each response was rated by three annotators, and the final score for each metric is the average of these three ratings. The inter-rater agreement, measured by Fleiss' kappa, is 0.43, indicating a moderate level of agreement among annotators.

From Table 5.3, our method performs the best among the methods. In addition, the seemingly small improvement is meaningful because the 0–5 scoring scale is highly constrained, and human annotators typically provide scores in a narrow range. This tendency to avoid extreme values means that any positive change, no matter how small, is a statistically significant indicator of model improvement. Aligned_{d-PM} obtained the highest score in all aspects, including the global consensus. It demonstrates that our method can generate less controversial and more helpful responses in the task of emotional support conversation.

5.3.3 Further Analyses

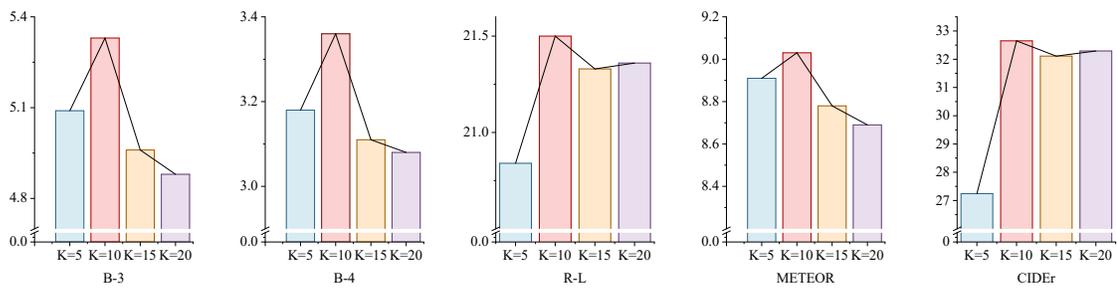


Figure 5.4: Alignment with d-PM: Model performances with different candidate numbers K .

The effect of candidate number K during calibration. To investigate how the number of candidates K affects model calibration, we adjusted the MultiESc calibration procedure by varying K among 5, 10, 15, and 20. This corresponds to modifying the beam width in the diverse beam search. In principle, increasing the number of candidates allows the model to consider more potential outputs, which could raise the theoretical upper bound of performance. However, as shown in Fig. 5.4, performance initially improves with larger K but eventually declines. This decrease occurs because excessively many candidates may include redundant or highly similar samples, making it difficult for the generation model to discriminate effectively between them.

Metric	MultiESC	RL	Ours
B-1	20.36	11.75	21.59
B-2	8.80	4.81	9.56
B-3	4.92	2.78	5.33
B-4	3.14	1.81	3.36
R-L	21.00	19.57	21.50
METEOR	8.58	6.25	9.03
CIDEr	30.69	26.02	32.65
Extreme	52.74	51.21	53.15
#(Samples)/s	-	2.65	4.36

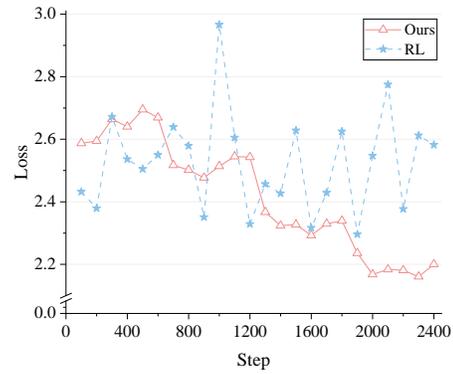


Figure 5.5: Alignment with d-PM: Comparison between alignment with RL and Ours.

Contrastive Learning vs. Reinforcement Learning. We adopt contrastive learning to align generation models with human preferences, instead of the commonly used reinforcement learning (RL) paradigm. This choice is motivated by several factors. First, RL typically requires costly online decoding, whereas contrastive learning can be performed offline in a single pass [165]. Second, RL often suffers from slow convergence [47]. To illustrate this, we also applied RL to align MultiESC. As shown in Fig. 5.5, RL processes fewer training samples per second compared to our contrastive learning framework. The left panel reports automatic evaluation results (#Samples/s denotes the number of samples trained per second), and the right panel shows the training loss over steps. Even after the same number of training steps, RL exhibits higher loss and substantially lower performance than our method.

Case study A case presenting responses generated by various models is shown in Table 5.4. **Green text:** the help-seeker’s emotion and problem. **Blue text:** the conversation context (history). **Orange text:** human evaluation results.

***** Case One *****

The Help-seeker's Emotion: Anxiety

The Help-seeker's Problem: Need a second job to pay off credit card bills.

Problem Type: Job crisis

Supporter: Hello, how are you today ?

Seeker: I'm good, busy. I need another job to pay more bills .

Base I'm sorry to hear that. Is there anything you'd like to talk about?

Overall 2,1,2

Global 2,2,2

Consensus

Aligned_{major} I'm sorry to hear that. Would you like to talk about it?

Overall 3,2,2

Global 3,3,2

Consensus

Aligned_{soft} I'm sorry to hear that. Would you like to talk about it?

Overall 3,2,2

Global 3,3,2

Consensus

Aligned_{d-PM} I'm sorry to hear that. Is there a way you can get a second job to pay off your debts?

Overall 4,5,4

Global 4,4,5

Consensus

***** Case Two *****

The Help-seeker's Emotion: Anxiety

The Help-seeker's Problem: Applying for work, I have been retired for 2 years and would like to return to the workforce.

Problem Type: Procrastination

Seeker: I have been retired for over 2 years and would like to return to the workforce, part-time.

Supporter: So you are thinking of going back to work?

Seeker: Yes, it would help us greatly.

Supporter: You have the right idea already.

Seeker: There are plenty of jobs and not sure where to start.

Base	That is a great idea. I think it would be great to get your mind off of retirement.		
Overall	2,3,3	Global	2,4,3
		Consensus	

Aligned_{major}	Yes, there are plenty of jobs out there. You just need to know where to start.		
Overall	4,4,3	Global	3,4,3
		Consensus	

Aligned_{soft}	I see. So you are thinking about getting back into the field?		
Overall	4,4,4	Global	4,4,4
		Consensus	

Aligned_{d-PM}	That is a very good question. Would you be open to looking for a new job?		
Overall	4,5,4	Global	4,4,4
		Consensus	

***** Case Three *****

The Help-seeker's Emotion: Sadness

The Help-seeker's Problem: I have a friend who knew me since I was little, she is closer to my mother's age so she is kinda like my aunt growing up. She always treats me like I'm little. But I'm a grown woman now and I want my relationship to go far and grow more.

Problem Type: Problems with friends

Seeker: I have a friend who knew me since I was little. She is a little older than me. She still treats me like I'm still little. I tried to explain to her that I'm a grown woman and I grew out of a lot of things when I was a kid.

Supporter: Is she listening to you? I know how it feels to be patronized by people who are older.

Seeker: She still didn't get it. I want my relationship to move further but with her like this. I feel sad. Is there anything I can do?

Supporter: Well, if she is not listening to you when you try to talk to her you have to ask yourself whether the relationship is really worth it?

Seeker: That's true . I didn't think about it. I know she is a good person. So I would give her a chance and I can talk about it before.

Base	I think it's always worth talking to someone who is older than you.		
Overall	2,1,3	Global	2,3,3
		Consensus	
Aligned_{major}	Well, it's always worth talking to someone who has been through a similar situation.		
Overall	3,4,2	Global	3,3,2
		Consensus	
Aligned_{soft}	Well, it's always worth talking to someone who has been through a similar situation.		
Overall	3,4,2	Global	3,3,2
		Consensus	
Aligned_{d-PM}	I think that's a great idea. I think it's great that you are willing to give it a try.		
Overall	4,5,5	Global	4,5,4
		Consensus	

Table 5.4: Alignment with d-PM: Case study.

5.4 Conclusion

In this chapter, we aim to align models with human preferences to finally advance the development of reliable emotional support conversational agents. Unlike prior approaches, inherent disagreements in human preferences are explicitly accounted for. This consideration is motivated by two key factors: first, achieving consensus is often unrealistic due to the highly subjective nature of emotional support; second, enforcing uniformity in preferences may unintentionally marginalize minority perspectives. To address these challenges, this chapter proposes a Bayesian framework—Preference Modeling with Disagree-

ment (d-PM)—which captures the nuances of disagreement from limited human feedback. The resulting preference scores are then used to calibrate existing text generation models. Experimental results on a benchmark emotional support dataset demonstrate the effectiveness of the proposed approach.

Chapter 6

Unhelpfulness Mitigation via Multifaceted AI Feedbacks

6.1 Introduction

This chapter deals with the second problem of reliable responding, which is how to mitigate unhelpful response generation. Previous emotional support conversational models occasionally generate well-intended responses that produce a counterproductive support effect, i.e., exacerbating the negative emotional states of users or inhibiting effective problem-solving, as shown in Fig. 6.1. In the psychology and communication theories, these failed support attempts are termed “*unhelpful messages*” [46, 17]. The frequency of unhelpful responses generated by some of the previous models is not extremely high, e.g, approximately 30% of responses generated by MultiESC [23] on ESConv benchmark [85] are identified as unhelpful under strict evaluation criteria. Unhelpful responses can undermine prior supportive interactions and erode the trust between a help-seeker and a supporter [88]. Thus, it is vital to reduce the chance that a model outputs such responses. We tackle this challenge through two complementary directions. *D1 – Incorporating Multiple Support Facets*: Many earlier studies focus on only one dimension of emotional support—such as em-

pathetic expression [78], conversational skill [23, 85], or overall coherence [34]. This single-focus design often overlooks other key aspects and may yield unhelpful replies [46], as shown in Fig. 6.1. *D2 – Explicit Suppression of Unhelpful Outputs*: Conventional models are usually trained by minimizing the negative log-likelihood of reference responses. To move beyond this objective, we aim to explicitly lower the probability of generating responses judged as unhelpful. To realize these goals, we introduce **Muffin**, a **mitigating-unhelpfulness** framework that leverages multifaceted AI feedback for emotional support. For *D1*, Muffin includes an AI feedback module that evaluates a candidate reply across multiple support dimensions. This module is obtained by instruction-tuning LLaMA, avoiding expensive manual annotation. For *D2*, the emotional-support model is optimized with a combined loss: the standard maximum-likelihood objective plus a contrastive term that directly penalizes responses flagged as unhelpful. These two components work together to reduce the production of unhelpful responses in emotional-support dialogues.

The main contributions of the work presented in this chapter are threefold:

- (1) We identify and address a critical challenge in current emotional support conversation models: their tendency to generate unhelpful responses, which undermines their reliability.
- (2) We propose Muffin, a novel, model-agnostic framework that uses a multifaceted AI feedback module to identify and mitigate unhelpful responses through contrastive learning.
- (3) We conduct extensive experiments on state-of-the-art emotional support models to demonstrate Muffin’s effectiveness in reducing the generation of unhelpful responses.

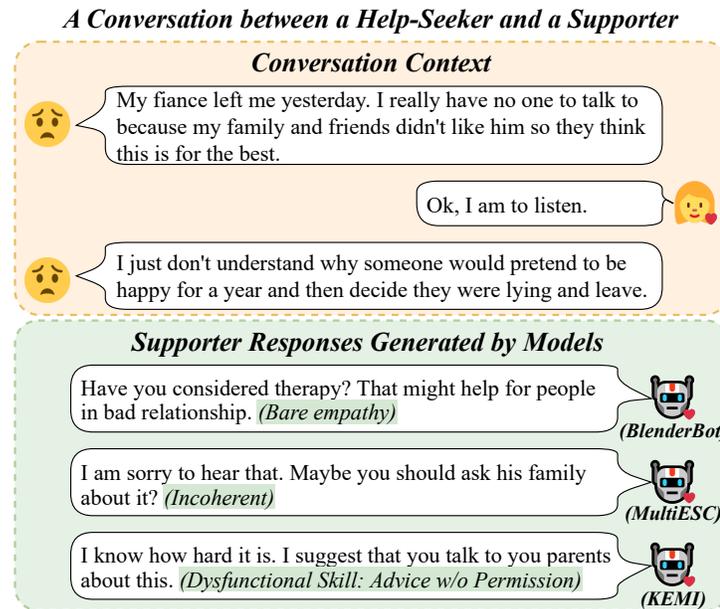


Figure 6.1: Muffin: Examples of unhelpful responses generated by various emotional support conversation models.

6.2 Method

6.2.1 Problem Formulation

We refer to established theories in psychology and communication [46, 17] and use the term “*unhelpful*” to describe responses that consistently result in negative outcomes during emotional support interactions. In contrast, responses that generate positive effects, or at minimum avoid causing harm, are labeled “*helpful*.” According to these theories, an unhelpful response often arises from a deficiency in a single facet of emotional support. Such a deficiency can intensify the user’s negative emotions or impede their problem-solving process. For example, a response may be classified as unhelpful if it either fails to acknowledge the user’s feelings and needs (demonstrating a lack of empathy) or criticizes the user’s behavior (reflecting a problematic communication skill such as confrontation). This characteristic serves as a key criterion for identifying un-

helpful responses.

The primary goal is to mitigate the generation of unhelpful responses. Rather than training a new model from scratch, we aim to refine a pre-trained emotional support conversation model with the dataset it was originally trained on. This process unfolds as follows. Let $\mathcal{G}(\theta_0)$ represent the model trained on a dataset \mathcal{D} , where θ_0 denotes the model parameters. Each instance in \mathcal{D} is denoted as (x, y) , with x as the input and y as the expected output. Usually, x is the conversation context, but it contains additional related information in some models. Assume that there are K samples $\{\hat{y}^1, \dots, \hat{y}^K\}$ with labels $\{\hat{l}^1, \dots, \hat{l}^K\}$. These samples are the diverse beam search generation results of $\mathcal{G}(x; \theta_0)$. As for the label $\hat{l}^k \in \{0, 1\}$, it represents feedback to indicate whether the sample \hat{y}^k is unhelpful ($\hat{l}^k = 0$) or not unhelpful ($\hat{l}^k = 1$). The objective is to refine the model's parameters θ such that the likelihood of generating unhelpful samples is reduced relative to helpful ones. In this process, we only modify the training process, ensuring that the model's architecture and the inference mechanism remain untouched. Moreover, the proposed approach is model-agnostic. This implies that $\mathcal{G}(\theta_0)$ can be any deep learning model designed and trained for emotional support conversations.

The overall framework of the proposed method, Muffin, is outlined in Fig. 6.2.  and  indicate helpful (non-unhelpful) and unhelpful labels, respectively. It is composed of two principal components, each specifically designed for deliberations as mentioned: **D1**: *Consideration of Multiple Facets* and **D2**: *Direct Minimization of Unhelpful Response Probability*, respectively. The multifaceted AI feedback module aims to identify whether a response from multiple facets of emotional support is unhelpful. The unhelpfulness mitigation module mitigates the likelihood of unhelpful responses by contrasting helpful and unhelpful re-

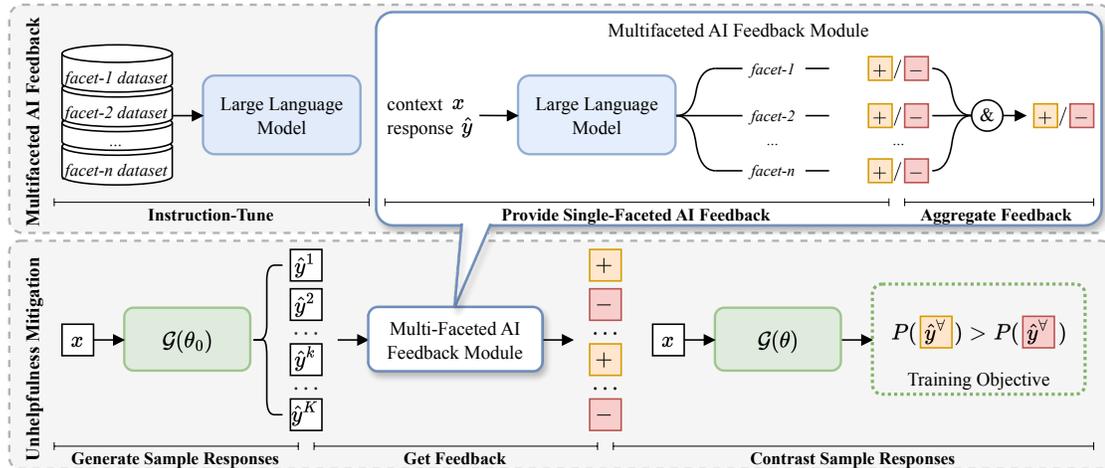


Figure 6.2: Muffin: Overview architecture.

sponses.

6.2.2 Multifaceted AI Feedback

We distinguish whether a response is unhelpful from multiple facets. However, collecting feedback from humans is inefficient and costly. In addition, recent large language models (LLMs), such as the GPT series [98] and LLaMA [127], demonstrate remarkable natural language understanding capabilities. Therefore, we decided to obtain feedback from AI.

Instruction-tuning Prompt engineering provides a simple and straightforward approach to obtaining feedback from LLMs. However, our experiments suggest that it is challenging to manifest the full potential of LLMs for emotional support without investing significant effort in prompt design, which will be detailed later. As an alternative, we elicit the desired capabilities of the LLM via instruction tuning [141]. Specifically, we design task descriptions and instructions tailored to classification tasks related to different emotional support facets. We use the relevant datasets for these tasks. The prompt format is shown in

Fig. 6.3. Importantly, the response class reflects whether a response is unhelpful for the given facet. During training, all *texts in italics enclosed in curly braces* are provided to the model. At inference time, the model is required to predict the *response class* based on the other *italicized inputs within the curly braces*.

```

### Instruction:
{task description and instruction}

### Input:
Conversation Context: {context}
The last supporter statement: {response}
{all possible classes}

### Output:
{response class}

```

Figure 6.3: Muffin: The prompt to obtain Multifaceted AI Feedback.

Multifaceted AI feedback module The final feedback for an emotional support response is derived from an aggregation of AI feedback across multiple facets. For each response along with its corresponding context, the instruction-tuned LLM is employed to evaluate all the predefined facets individually. If any facet indicates that the response is unhelpful, the response is assigned a label of 0; otherwise, it is considered helpful (non-unhelpful) and labeled as 1.

6.2.3 Unhelpfulness Mitigation

We mitigate $\mathcal{G}(\theta_0)$ to generate unhelpful responses by contrasting helpful and unhelpful responses generated by $\mathcal{G}(\theta_0)$ itself, which can be implemented by the following three steps:

Generating sample responses We utilize $\mathcal{G}(\theta_0)$ to generate responses on its own training dataset \mathcal{D} using diverse beam search [131]. Thus, for each instance

$(x, y) \in \mathcal{D}$, there are K sample responses $\{\hat{y}^1, \dots, \hat{y}^K\}$.

Getting feedback These responses can be generated because they have relatively high generation probabilities. However, some of them can be unhelpful responses. Therefore, we adopt the multifaceted AI feedback module to identify whether these responses are unhelpful. Thus, we obtain K labels $\{\hat{l}^1, \dots, \hat{l}^K\}$, where $\hat{l}^k \in \{0, 1\}$.

Contrasting sample responses We expect that the model \mathcal{G} can assign a higher likelihood to the helpful responses than the unhelpful ones. Therefore, we contrast them using the following loss:

$$\begin{aligned} \mathcal{L}_{cl} = & \frac{1}{2K} \sum_i \sum_{j \neq i} \max(0, \\ & -(\hat{l}^i - \hat{l}^j) \times (\text{P}(\hat{y}^i|x) - \text{P}(\hat{y}^j|x) + \lambda)), \end{aligned} \quad (6.1)$$

where λ is the margin hyperparameter. Moreover, $\text{P}(\hat{y}^i|x)$ is the length-normalized log-probability of the response \hat{y}^i , and it is computed by:

$$\text{P}(\hat{y}^i|x) = \sum_{t=1}^{|\hat{y}^i|} \frac{\log \mathcal{G}(\hat{y}_t^i|x, \hat{y}_{<t}^i; \theta)}{|\hat{y}^i|^\alpha}, \quad (6.2)$$

where α is the length penalty hyperparameter. In addition to the above loss, we also consider the negative log-likelihood loss to prevent the model's generation from deviating too much from the ground truth. The loss can be formulated as:

$$\mathcal{L}_{gen} = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log \mathcal{G}(y_t|x, y_{<t}; \theta). \quad (6.3)$$

The final loss is the combination of the above two losses:

$$\mathcal{L} = \beta_{cl} \mathcal{L}_{cl} + \beta_{gen} \mathcal{L}_{gen}, \quad (6.4)$$

where β_{cl} and β_{gen} are weight hyperparameters.

6.3 Experiments

6.3.1 Experimental Settings

Base models ($\mathcal{G}(\theta_0)$) The proposed method, i.e., Muffin, is a model-agnostic approach designed to mitigate the unhelpfulness of an existing emotional support conversation model. To examine its effectiveness, we experiment with five recent models: Vanilla, Joint, MultiESC, TransESC [164], and KEMI [34]. The first three models have been introduced in Section 5.3.1. The last two models are detailed as follows:

- KEMI incorporates various knowledge for a mixed-initiative conversation model, which can provide emotional support [34]. Consequently, the response coherence is also enhanced.
- TransESC [164] predicts the transitions of the user’s emotion, the communication skills, and the conversation keywords. Then, such information is used for response generation. This model takes into account more than one facet of emotional support; however, it is not optimized to reduce the likelihood of unhelpful responses like other models.

We obtain each model’s parameters θ using its official implementation and the default hyperparameters.

During the training of $\text{Muffin}_{\mathcal{G}(\theta_0)}$, we adopt the same hyperparameters as the base model $\mathcal{G}(\theta_0)$, such as batch size and random seed. To facilitate more stable convergence, a smaller learning rate of 3×10^{-5} is applied. The training is performed for a single epoch, as the loss stabilizes within one pass over the data. Other hyperparameters are set as follows: margin $\lambda = 0.01$, length penalty $\alpha = 1$, and weights $\beta_{cl} = \beta_{gen} = 1$. For response generation, as outlined in

Section 6.2.3, we sample $K = 10$ candidate responses. Accordingly, both the beam size and the number of beam groups are set to 10, while all remaining generation settings follow those of the base model.

ESConv dataset (\mathcal{D}) The ESConv dataset [85] is used to train the aforementioned base models. All base models, with the exception of TransESC, follow the original division of ESConv for training, validation, and testing, using an 8:1:1 ratio. TransESC employs a random split while maintaining the same ratio. Notably, each model adopts different data preprocessing methods. For consistency, we adopt the original data division and preprocessing protocols of each base model.

Facets of Emotional Support Considered in Muffin and Dataset We consider three essential facets of emotional support: empathetic expression, skill efficiency, and response coherence, which the base models incorporate into their models. Here, we would like to describe the unhelpfulness of each facet and detail the corresponding classification dataset.

Empathetic expression Meaning: Empathetic expressions reflect the supporter’s understanding and engagement with the help-seeker’s perspective. Their absence can hinder conversation flow and trust-building [94]. While empathy is multifaceted [136, 99, 12], we adopt the framework by [118], which defines three mechanisms: emotional reactions, interpretations, and explorations. Each is rated as no, weak, or strong communication. Responses lacking empathy across all mechanisms are deemed unhelpful. For example, “Sleeplessness can result in upsetness,” fails to convey empathy when responding to “I am upset.” **Dataset:** We use the dataset from [118], containing 3K context-response pairs annotated under this frame-

work, to train and evaluate models in detecting unhelpful empathetic responses. Responses consistently labeled as “no communication” across all these mechanisms are identified as unhelpful.

Skill efficiency Meaning: Effective communication strategies enable supporters to deliver appropriate and impactful messages [46], enhancing their understanding of the help-seeker and facilitating problem-solving [52]. However, some dysfunctional skills can have adverse effects [8, 18]. For instance, while advice is generally helpful, unsolicited advice may be less effective than offering no specific support at all. **Dataset:** This study evaluates skill efficiency using motivational interviewing (MI) skill codes [95], which classify responses into three categories: MI adherent, MI non-adherent, and others. Responses labeled as MI non-adherent are considered unhelpful. To train and evaluate models, we use the MI dataset by [144], which includes 17K context-response pairs annotated with the three MI codes. Responses labeled as “MI non-adherent” are considered unhelpful.

Response coherence Meaning: While response coherence is a fundamental expectation in almost all conversational systems [55], it holds particular significance in emotional support conversation systems. Incoherent responses can disrupt communication and cause misunderstanding. To mitigate this issue, we classify responses into two categories: coherent and incoherent, with the latter considered unhelpful. **Dataset:** We constructed a dataset specifically for detecting response coherence, based on the base model’s training set \mathcal{D} . In particular, 4K context-response pairs were randomly sampled from \mathcal{D} , with the original responses labeled as coherent. To generate incoherent responses, we employed two strategies: (1) replacing the original response with one from a different conversation, and (2) altering

key words or critical information in the original response to produce a subtly incoherent variant. Using these procedures, we obtained around 12K context-response pairs in total, which are used to train models for identifying unhelpful responses with respect to coherence.

6.3.2 Evaluation

Multifaceted AI feedback For the multifaceted AI feedback module, a 7B LLaMA model [127] is instruction-tuned using a low-rank adapter (LoRA) approach [54]. During tuning, the LLaMA weights are frozen, and trainable rank decomposition matrices are inserted into the query, key, value, and output layers. The training data comprises 22K instances, formed by merging the datasets described in Section 6.3.1, with each context-response pair formatted according to the instruction template in Fig. 6.3. The dataset is partitioned into training, validation, and test sets following an 8:1:1 split ratio. The learning rate is 3×10^{-4} , and the training epoch is 12.

The Multifaceted AI Feedback Module can be implemented with various backbone models. As shown in Fig. 6.4, the module's performance with different backbones, including GPT-3.5, GPT-4, vanilla LLaMA, and instruction-tuned LLaMA, is evaluated on the test set. The performance is compared across different facets of emotional support, as well as the aggregated feedback. Accuracy and F1 scores are reported to reflect model performance across the entire test set and for each individual facet. Without any fine-tuning, GPT-4 demonstrates superior classification performance compared to the other two models. A detailed analysis reveals that GPT-3.5 tends to label responses as non-unhelpful, while the vanilla LLaMA model often assigns responses to a single category across multiple facets, resulting in the poorest performance. In con-

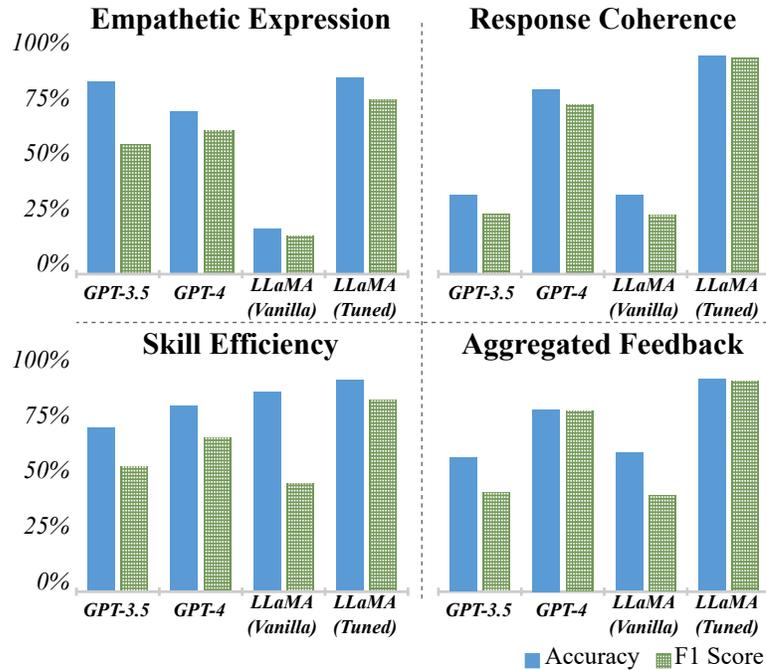


Figure 6.4: Muffin: Multifaceted AI feedback performances.

trast, instruction-tuning substantially improves LLaMA, achieving an accuracy of 90.72% and an F1 score of 89.86% on the aggregated feedback, surpassing the other models. This significant enhancement supports our choice of employing the 7B instruction-tuned LLaMA in the Multifaceted AI Feedback Module.

Automatic Evaluation of Muffin The automatic evaluation of Muffin includes the following metrics: BLEU, ROUGE, METEOR, CIDEr, and Extreme (a BOW Embedding-based matching score). Table 6.1 showcases the performance of Muffin with different base models in all automatic evaluation metrics. For all metric scores and feedback, a higher value indicates better performance. The values marked with * indicate the results are statistically significant with $p < 0.05$. In general, Muffin demonstrates significant enhancements across nearly all evaluation metrics. Moreover, it can be observed that the performance of $\text{Muffin}_{\mathcal{G}(\theta_0)}$ is predominantly influenced by its base model $\mathcal{G}(\theta_0)$, assessed

Model	B-1	B-2	B-3	B-4	R-L	METEOR	CIDEr	Extreme
Vanilla	18.23	7.02	3.49	1.99	16.09	7.31	14.95	50.48
Muffin _{Vanilla}	19.43*	7.58*	3.66	2.02	16.26	7.72*	13.90*	51.00*
Joint	18.77	7.54	3.79	2.15	17.72	7.59	17.38	50.96
Muffin _{Joint}	20.59*	8.38*	4.26*	2.54*	18.35*	8.18*	19.12*	51.46*
TransESC	17.32	7.10	3.63	2.18	17.47	7.53	22.07	51.33
Muffin _{TransESC}	17.19	7.17*	3.73*	2.25*	17.54*	7.58*	22.72*	51.57*
KEMI	19.85	8.15	4.24	2.52	17.17	7.92	15.09	50.85
Muffin _{KEMI}	20.01*	8.31*	4.36*	2.60*	17.30*	7.99*	15.45*	51.11*
MultiESC	21.79	9.19	4.98	3.05	20.92	8.93	28.84	52.59
Muffin _{MultiESC}	21.83*	9.28*	5.12*	3.21*	21.26*	8.92	31.26*	52.83

Table 6.1: Muffin: Automatic evaluation results.

Here is a conversation between a help-seeker and a support. Imagine that you are the help-seeker, and compare the following two responses. If you find both responses to be equally effective or unsatisfactory, please indicate your assessment as a "Tie".

Conversation History:

Help-seeker: Hi i am okay, a little bit sad though.

Support: Okay. I am very sorry to hear that! Do you want to tell me more about that?

Help-seeker: Well with the holidays coming up I have been very stressed and nervous about what i am going to do.

Two supporter responses:

A. Of course! I am sorry to hear you are stressed about the holidays. Can I ask what are you worried about?

B. Of course! I am sorry to hear you are feeling stressed and anxious about the holidays.

1. Which response is more fluent (grammar errors and inappropriate repetition can decrease the fluency)? A, B or Tie?

>

2. Which response is more skillful in identifying the help-seeker's problem? A, B or Tie?

>

3. Which response is better at comforting the help-seeker? A, B or Tie?

>

4. Which response can give more helpful and information suggestions? A, B or Tie?

>

5. Which response is generally more helpful? A, B or Tie?

>

Figure 6.5: Muffin: An example presented to annotators for evaluation.

through automatic evaluations.

Human Evaluation of Muffin Following previous work [34, 85], five aspects are considered in human evaluation, and they are: (1) *Fluency*: which model's response is more fluent? (2) *Identification*: which model's response is more skillful in identifying the user's problem? (3) *Comforting*: which model's response is better at comforting the user? (4) *Suggestion*: which model can give more helpful and informative suggestions? (5) *Helpfulness*: which model's response is gener-

ally more helpful from the aspect of the help-seeker? For each $\mathcal{G}(\theta_0)$ -Muffin $_{\mathcal{G}(\theta_0)}$ pair, 100 instances are randomly selected for comparison. Four unique human evaluators are asked to vote on which response is better, with the option to select “tie” if both responses are considered equally effective. The final results are averaged from their evaluations. To avoid bias from the order in which responses are presented, the sequence of the two responses is randomized for each evaluation. We asked four PhD students who have knowledge about mental health support to participate in the human evaluation. Prior to the evaluation, annotators are briefed on the characteristics of emotional support conversations to ensure they understand the goals of the task. During the assessment, they are instructed to place themselves in the role of the help-seeker. Given the conversation context, annotators compare two generated responses, as illustrated in Fig. 6.5. To safeguard their well-being, annotators are allowed to pause or stop if they encounter content that makes them uncomfortable. Compensation is provided at 1.5 to 2 times the local minimum hourly wage. On average, annotators spend approximately 40 seconds per response pair. The inter-rater agreement, measured by Fleiss’s Kappa, is 0.39, which is relatively high for subjective evaluations, typically ranging from 0.2 to 0.6 [149, 28]. For aggregation, the winning response is determined by annotator consensus: if two annotators prefer response A and the remaining two indicate a tie, response A is selected as the winner; an equal split in preference results in a tie label.

Fig. 6.6 summarizes the A/B test results on BlenderBot-Joint [85], KEMI [34], and MultiESC [23], along with their corresponding Muffin models. Displayed within each bar, from left to right, are the ratios for “Muffin Wins,” “Tie,” and “Base Wins.” These three settings are selected for their significant performance in automatic evaluation. The inter-rater agreement, i.e., Fleiss’ Kappa [40], is

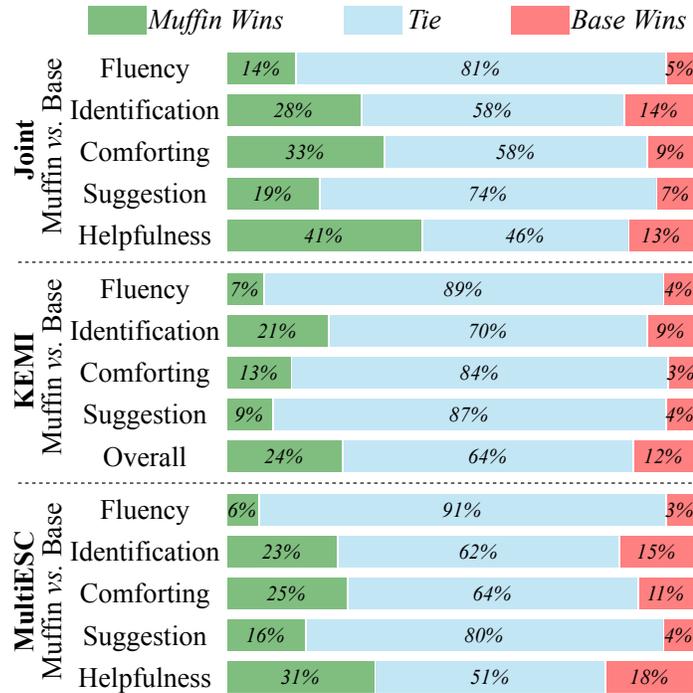


Figure 6.6: Muffin: Human A/B test results.

0.39, implying fair agreement. The Muffin models are regarded as more helpful in general, as evidenced by their higher *Helpfulness*. Responses generated by Muffin models are slightly more fluent than those generated by base models. It is also observed that ‘ties’ are common in evaluations of response fluency, mainly because the responses generated are typically fluent. Compared with the corresponding base model $\mathcal{G}(\theta_0)$, the $\text{Muffin}_{\mathcal{G}(\theta_0)}$ model shows some more powerful capability in identifying the help-seeker’s problem. Moreover, $\text{Muffin}_{\mathcal{G}(\theta_0)}$ models can generate responses that have better effects to comfort the users than $\mathcal{G}(\theta_0)$. Annotators also prefer responses generated by $\text{Muffin}_{\mathcal{G}(\theta_0)}$ because of their more helpful and informative suggestions. These results prove that Muffin indeed mitigates the unhelpfulness of emotional support conversation models.

Model	Empathetic Expression	Skill Efficiency	Response Coherence	Aggregated Feedback
<i>A. Generating One Response</i>				
Vanilla	81.22	90.43	80.69	64.83
Muffin _{Vanilla}	85.83	92.82	84.33	71.26
Joint	80.61	88.93	80.65	61.48
Muffin _{Joint}	82.33	90.47	83.04	64.76
TransESC	81.06	91.44	74.76	63.24
Muffin _{TransESC}	81.28	91.49	78.67	66.02
KEMI	83.01	88.47	85.76	70.90
Muffin _{KEMI}	83.40	88.68	87.15	72.33
MultiESC	83.39	90.66	85.38	70.00
Muffin _{MultiESC}	85.24	92.92	86.06	72.38
<i>B. Generating Ten Responses</i>				
Vanilla	81.90	92.03	78.89	64.51
Muffin _{Vanilla}	89.84	96.05	82.60	73.78
Joint	79.31	90.56	79.75	63.30
Muffin _{Joint}	83.43	92.26	80.02	66.54
TransESC	74.16	92.78	43.93	41.21
Muffin _{TransESC}	74.20	92.92	44.15	41.43
KEMI	81.16	87.49	81.57	66.23
Muffin _{KEMI}	81.39	87.64	81.61	66.47
MultiESC	78.61	89.67	77.33	60.83
Muffin _{MultiESC}	78.77	89.73	77.26	60.87

Table 6.2: Muffin: Multifaceted AI feedback results.

Helpfulness Evaluation via AI Feedback of Muffin Table 6.2 provided the multifaceted AI feedback results as a reference. The AI feedback is sourced from the comprehensive AI feedback module. The top subtable showcases the percentage of each model’s helpful (non-unhelpful) responses employing the decoding strategy of the base model; the bottom subtable displays the percentage of each model’s helpful responses utilizing diverse beam search with a beam size set at 10. All values are expressed in percentages (%), where higher percentages signify superior performance. The values in the left subtable are statistically significant with $p < 0.05$.

We utilize the multifaceted AI feedback module to identify various models’ helpful (non-unhelpful) responses and compute their percentage, displayed in the left subtable. Furthermore, we analyze the helpful response percentage when each model generates ten responses using diverse beam search, reported

in the right subtable.¹

Overall, Muffin demonstrates enhancements in AI feedback across multiple facets. However, three intriguing phenomena emerge.

1. While the left subtable shows an increase in the frequency of helpful responses due to the Muffin framework, the evidence presented in the right subtable (Aggregated Feedback) is weaker, particularly when the base model is TransESC, KEMI, or MultiESC. This finding is consistent with our loss function (Eq. (6.1) and Eq. (6.4)). The contrastive loss is not designed to suppress unhelpfulness directly but rather to assign higher generation probabilities to helpful (non-unhelpful) responses. Consequently, when generating a single output, the model selects the one with the highest probability.
2. Despite lacking a dedicated mechanism for incorporating communication skills, **BlenderBot-Vanilla** attains a notably high score in Skill Efficiency. Upon closer examination, we observe that this model frequently produces responses like “I can understand that...” or “I’ve experienced something similar...” These responses, categorized as self-disclosure, do not undermine previous emotional support efforts and are not detected as “unhelpful” by the AI feedback module. However, a conversation dominated by self-disclosure may not fully align with the objective of providing genuine support.
3. Another noteworthy observation is that **TransESC** receives exceptionally low AI feedback scores for Response Coherence, even though it incorporates external knowledge. This insight aligns with findings from Trans-

¹The assertion “approximately 30% of responses generated by MultiESC on the ESConv benchmark are identified as unhelpful” is derived from the findings presented in this table.

Model	B-1	B-2	B-3	B-4	R-L	METEOR	CIDEr	Extreme
Joint	18.77	7.54	3.79	2.15	17.72	7.59	17.38	50.96
Muffin _{Joint}	20.59	8.38	4.26	2.54	18.35	8.18	19.12	51.46
Muffin _{Joint} (Empathetic Expression)	19.58	8.06	4.04	2.33	18.46	7.87	19.61	51.43
Muffin _{Joint} (Skill Efficiency)	18.68	7.51	3.81	2.19	17.98	7.67	17.84	50.98
Muffin _{Joint} (Response Coherence)	20.04	8.10	4.04	2.26	18.24	7.95	18.07	51.31

Table 6.3: Muffin: Ablation study results.

ESC’s ablation experiment [164], which suggests that the generated knowledge may contain noise, potentially explaining the low coherence scores.

6.3.3 Further Analyses

Ablation study To assess the impact of different feedback facets, an ablation study is conducted. Specifically, AI feedback from each individual facet is used to differentiate helpful from unhelpful responses, which are then contrasted to mitigate unhelpfulness. This study utilizes BlenderBot-Joint as the base model, chosen for its simple architecture and the significant improvements demonstrated by Muffin_{Joint}.

From Table 6.3, we can find that relying solely on AI feedback from a single facet for unhelpfulness mitigation results in diminished performance in automatic evaluations when compared to the comprehensive Muffin_{Joint} model. Muffin_{Joint} model. Muffin_{Joint} (X) indicates the mitigation process only uses AI feedback in terms of the facet X. All values are statistically significant with $p < 0.05$. This underscores that considering multiple facets of emotional support when building emotional support conversation models is necessary.

Another insight from Table 6.3 is that all three ablated models outperform the base model, suggesting that the proposed strategy of directly mitigating unhelpful responses (D2) is both reasonable and effective. The results also high-

<i>Conversation History</i>	
Help-seeker	: I'm an alcoholic. I get that thought in my head and its like I become obsessed.
Supporter	: Why not try to put attention on some other things instead of a beer?
Help-seeker	: Hmm like what?
<i>Model Generated Responses</i>	
Joint	You can try to be normal. That is a great idea. (<i>Lacking empathy</i>)
Muffin _{Joint}	Yes, there are some things that you can do, but you have to be aware of your own mental health, that's why it's important to know you are doing the right thing.
MultiESC	It sounds like you have a lot on your plate and you want to be normal. (<i>Inconsistent</i>) I can understand how that would make you feel.
Muffin _{MultiESC}	I think you should focus on yourself instead of a beer. Do you have any hobbies that you enjoy?
KEMI	I think you should try to find some new things to do to relax.
Muffin _{TransESC}	Have you tried meditation? It can help you relax and relax.

Table 6.4: Muffin: Case study.

light that the quality of helpful and unhelpful examples plays a critical role in the success of unhelpfulness mitigation. Additionally, the impact of each facet on overall performance varies, which is both expected and reasonable. By examining the ablation results in Table 6.3 alongside the facet-specific findings in Table 6.2, different proportions of non-unhelpful responses across facets are observed. Notably, the skill efficiency facet shows the lowest proportion, explaining the weaker performance of Muffin (skill). In contrast, facets such as empathy and coherence show more balanced distributions, resulting in comparable performance for Muffin (Empathetic Expression) and Muffin (Response Coherence). These observations underscore the nuanced, facet-dependent effects on the efficacy of the proposed method.

Case study To intuitively illustrate the superiority of Muffin over its base models, a comparative case study is presented in Table 6.4, showcasing responses generated by three prominent base models (BlenderBot-Joint, MultiESC, and KEMI) and their corresponding Muffin versions. From the comparison of re-

sponses from BlenderBot-Joint and Muffin_{Joint}, we observe that the former tends to state facts directly, subtly implying that the help-seeker might be in an abnormal state. Such a statement ignores the help-seeker’s feelings and expresses minimal empathy. In contrast, Muffin_{Joint} conveys concern for the help-seeker’s well-being and attempts to solve the problem by shifting the help-seeker’s perspective, thereby amplifying the empathetic undertone. In the case of Multi-ESC, Muffin_{MultiESC} crafts a response that aligns more closely with the context, effectively addressing the inconsistency present in the response generated by its base model. Lastly, when comparing KEMI with Muffin_{KEMI}, Muffin’s version stands out as more beneficial. While KEMI’s response does not exhibit glaring issues, Muffin_{KEMI} offers a more specific and actionable recommendation, closely aligning with the help-seeker’s request for precise advice, in contrast to KEMI’s general guidance.

6.4 Conclusion

This chapter addresses the critical problem of unhelpful responses in emotional support conversational agents. We introduce Muffin, a novel, model-agnostic framework designed to mitigate this issue. Muffin uses a multifaceted AI feedback module to discern between helpful and unhelpful responses. By contrasting these outputs, our framework effectively reduces the generation of unhelpful content. Experiments on both automatic and human evaluations confirm Muffin’s effectiveness, showcasing its ability to significantly enhance the reliability of emotional support conversations.

Part III

Engaging Interaction

While the previous two parts have explored empathetic understanding and reliable responding of conversational agents, these capabilities alone do not ensure the overall effectiveness of a conversation, particularly in terms of fostering emotional engagement. Investigating the overall effectiveness of conversational agents poses a significant challenge, as it requires user involvement, which can lead to difficulties in evaluating and improving interactions. Prior research has attempted to address this challenge by analyzing strategic dialogue flow and linguistic choices, such as wording and dialogue acts [24, 23]. These approaches are based on the assumption that incorporating relevant knowledge and employing appropriate conversational strategies can enhance user engagement. However, the relationship between such strategies and user engagement is nuanced and not always straightforward. For instance, while well-crafted questions or informative content may encourage users to participate more actively, poorly phrased questions or excessive information can lead to disengagement [22]. Works in this part address these challenges with a focus on evaluating the emotional engagement level of interactions with the conversation agents. One of the most direct and meaningful indicators of user engagement is the user's reaction following the conversation. Whether an interaction achieves its intended purpose often depends on how the user feels about the exchange. Therefore, when investigating the effectiveness of conversational agents, it is essential to prioritize the user's perspective. To address this, this part introduces the use of simulated users for evaluating emotionally intelligent conversational agents. In addition, conversation-level features are paid attention to in these studies.

Chapter 7

A Client-Centric Automatic Evaluation of LLM Therapists

7.1 Introduction

Ever since ELIZA, a therapy chatbot, was found to provide emotional support, it has been argued that chatbots could scale up mental health support [27]. The recent advancements in large language models (LLMs) have bolstered the argument for their use as therapeutic tools, as evidenced by both research [97, 49] and reports from end-users [101]. Although many users report that LLM-based therapists provide helpful support [110], potential risks and harms have also been noted [31, 153]. Prior research has explored the capabilities and limitations of LLMs in therapeutic roles by evaluating their behavior [24, 72]. Nonetheless, these evaluations mainly consider the therapist’s perspective, which can differ considerably from that of the client [72, 35, 3]. In contrast, our work focuses on examining LLM therapists from the viewpoint of the client.

For a client-centered assessment, involving clients is critical. In clinical medical education, “actors” are hired and trained as simulated clients to interact with therapists for assessment purposes [69]. However, using this approach to assess LLM therapists presents significant challenges, particularly regarding scalabil-

ity. Beyond the costs of hiring and training, this method introduces ethical and technical hurdles. Ethically, the long-term simulation of client symptoms can cause discomfort for individuals [14], and this approach risks exposing them to potentially harmful LLM outputs. Technically, human behavior varies over time and across interactions, which makes it difficult to consistently compare the performance of different LLM therapists with the same client.

As an alternative, we involve clients simulated by LLMs, which demonstrate great potential in role-playing [20], in the assessment. An LLM client, equipped with a specific psychological profile, interacts with an LLM therapist, i.e., the evaluation target. After the interaction, the simulated client completes questionnaires, including items that ask the client about their feelings and thoughts in the counseling session. The client-centered assessment of the LLM therapist is derived from the results of these questionnaires. The entire procedure is illustrated in Fig. 7.1, and we refer to this framework as **ClientCAST**, which stands for a Client-Centered method for Assessing LLM Therapists through client simulation.

To prove the reliability of ClientCAST, we have conducted experiments using two human-human therapy counseling datasets, High-Low Quality Counseling [105] and AnnoMI [151], and four LLMs, i.e., Claude-3, GPT-3.5, Mixtral 8×7B and Llama 3-70B, to simulate clients. The findings indicate that simulated clients consistently adhere to the provided problems, reasons for visiting, symptoms, and speech tone, but they are less accurate in reproducing defined apparent traits. In general, more advanced models yield more accurate simulations. Additionally, high- and low-quality counseling sessions can be distinguished based on the completed questionnaires. Finally, we use ClientCAST to evaluate therapists implemented by Claude-3, GPT-3.5, Mixtral 8×7B, and Llama 3-70B.

The main contributions of this work are summarized as follows:

- (1) We introduce ClientCAST, a novel framework for evaluating LLM-based therapists from the perspective of clients by incorporating LLM-simulated clients into the assessment process.
- (2) Experimental results indicate that the simulated clients largely adhere to their assigned psychological profiles and are capable of distinguishing between high- and low-quality counseling sessions.
- (3) Leveraging ClientCAST, we systematically evaluate the performance of therapists implemented using various LLMs.

7.2 Method

7.2.1 Problem Formulation

To enable the automatic evaluation of LLM therapists, LLMs are used to simulate clients. Then, the feedback from the simulated clients obtained via questionnaires is used to compute three metrics: session outcome, therapeutic alliance, and self-reported feelings, which can reflect the LLM-therapist's performance. Fig. 7.1 presents an overview of the ClientCAST framework. It utilizes an LLM as a simulated client equipped with a specific psychological profile. The simulated client interacts with an LLM therapist and completes questionnaires regarding their interaction. Finally, ClientCAST provides a client-centered assessment of the LLM therapist based on the results of questionnaires.

To establish the reliability of ClientCAST, it is essential to determine its ability to evaluate therapy sessions in a manner consistent with human judgment. This can be achieved by deploying ClientCAST to assess existing therapy ses-

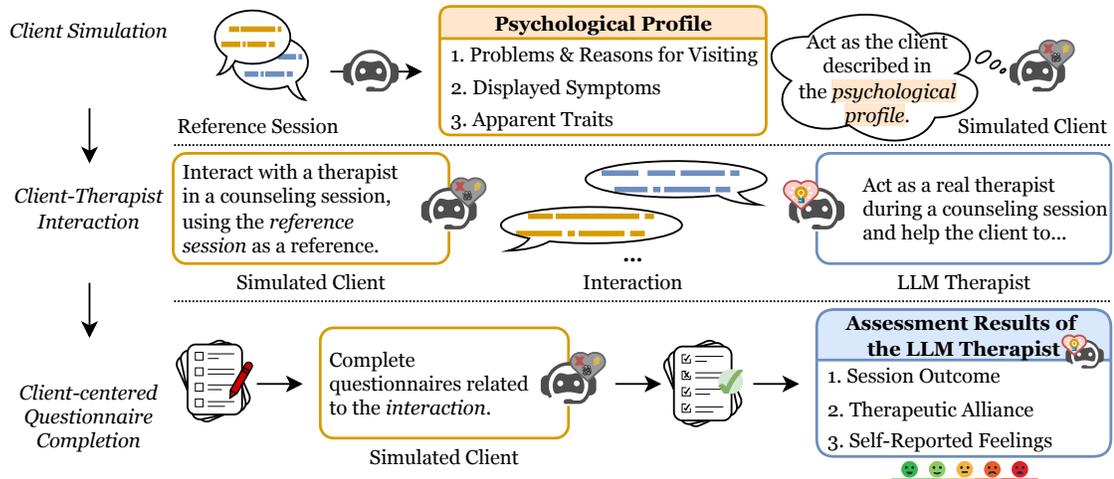


Figure 7.1: ClientCAST: Overview Architecture.

sions and then comparing its evaluation outcomes against human-assigned labels. In addition, to ensure the process is correct, the evaluation of the simulated clients is also conducted.

7.2.2 Client Simulation

Psychological Profile

The psychological profile provided to LLMs should significantly influence the counseling session and facilitate the identification of the specific client. Inspired by [115] and [21], we consider the following information: **(1) Problems & Reasons for Visiting.** We use two sentences to describe the client’s problems and the reasons for visiting, respectively. **(2) Displayed Symptoms.** We consider 61 potential client symptoms, covering areas of depression, anxiety, symptom distress, interpersonal relations, and social roles. These symptoms are from three widely used questionnaires: PHQ-9, GAD-7, and OQ-45. **(3) Apparent Traits.** Apparent traits significantly affect the client’s speech tone and conversation engagement. We account for the big five personality traits, emotion fluctu-

ations (Fluctuations), unwillingness to express emotions (Unwillingness), and resistance toward the therapist (Resistance). Each trait can be described with a severity level. There are five levels for the big five and three for the other three traits. A psychological profile example is in Section [7.2.2](#).

Simulation Method

We formulate the client simulation as follows. Let S_i represent the transcript of a counseling session involving a therapist T_i and a client C_i . The objective is to simulate C_i by prompting an LLM. First, we extract the psychological profile $\mathcal{P}_C(S_i)$ of C_i from S_i using the LLM. The LLM is then instructed to behave like C_i based on $\mathcal{P}_C(S_i)$. The session S_i is also supplied as a reference session. The LLM is tasked with learning the client's speech tone and conversational style from S_i but engaging in a new session with a different therapist, as if in a parallel universe. Prompts are shown as follows:

Here is a patient simulation task for you: act as a real patient ([Client]) during a counseling session to assess the psychotherapist's ([Therapist]'s) capabilities.

CASE SYNOPSIS: problems, name, gender, age and occupation

REASONS FOR TODAY'S VISIT: {reasons for visiting}

SYMPTOMS YOU ARE EXPERIENCING: {symptoms} You must manifest all the above symptoms through interaction.

YOUR APPEARANCE DURING ENCOUNTER: {apparent traits}

AN CONVERSATION BETWEEN THE PATIENT TO SIMULATE AND ANOTHER THERAPIST: {reference counseling session} While the context of the previous conversation should not influence this new session between [Therapist] and you, closely mimicking [Original Client]'s communication style, including the tone, word choice, sentence structure, expression of feelings, symptoms or issues. You can adopt the exact expressions/responses used by [Original Client] in the above conversation if necessary.

ADDITIONAL GUIDELINES: As a standardized patient to contribute to the assessment, adhere strictly to the provided guidelines and embody the described personalities or characters. To simulate a real patient effectively, you should: (1). Communicate your problems and feelings in a vague and colloquial manner. (2). Keep your responses very short. (3). Limit the amount of information in each response to ensure you have content for future conversations; instead, reveal details gradually through interaction. (4). If you find it's not the right moment to elaborate, a brief acknowledgment like 'I see' or 'okay' will suffice.

7.2.3 Client-centered Questionnaire Completion

Assessment Aspects & Questionnaires

Inspired by the work of [90], our assessment framework focuses on three key aspects of the counseling session:

1. **Session Outcome:** This evaluates the immediate impact of the session on the client's progress toward their therapeutic goals and their perception of the session's overall effectiveness.
2. **Therapeutic Alliance:** This assesses the quality of the client-therapist relationship, with a focus on the client's level of trust and agreement with the therapist.

3. **Self-Reported Feelings:** This captures the client’s immediate emotions, providing a subjective perspective on their feelings about themselves and the session.

To measure these aspects, we consider four emotional dimensions: *Depth*, *Positivity*, *Smoothness*, and *Arousal*. These dimensions, particularly *Depth*, are closely related to the perceived helpfulness of a session [29, 7]. Additionally, *Arousal* is known to have a strong correlation with a client’s ambivalence toward change [116, 37]. Correspondingly, we utilized five commonly used clinical psychology questionnaires: SRS, CECS, SEQ, WAI-SR, and HAQ-II.

Completion & Assessment Results

We prompt the LLM to complete a series of questionnaires. The prompt includes the client’s problems and reasons for seeking therapy, their apparent traits, and a transcript of their interaction with the LLM therapist. Each questionnaire item is presented with its rating scale and a clear explanation of what each scale value represents. The final assessment score for each aspect is calculated based on the relevant item scores, with higher values indicating better performance.

7.3 Experiments

In this section, ClientCAST is used to evaluate therapists implemented by different LLMs after the examination of its reliability.

7.3.1 Experimental Settings

Datasets Sessions are drawn from two datasets: **High-Low Quality Counseling** [105] and **AnnoMI** [151]. Both contain counseling therapy transcripts cu-

rated from publicly available videos on platforms such as YouTube and Vimeo. These transcripts uphold high-quality standards while protecting sensitive personal information. Each session is evaluated and labeled as either high- or low-quality based on MI psychotherapy principles. The datasets are used in accordance with the licensing and usage terms provided by their authors. These two datasets are employed both to simulate clients and to validate the client simulation. Transcripts lacking sufficient conversational context (e.g., return visits with minimal background) or those with limited dialogue (e.g., only a few turns) are excluded. In total, 213 high-quality and 87 low-quality sessions are included. Most conversations focus on clients' positive behavioral changes—for example, 21.67% address reducing alcohol consumption, 17.67% focus on weight management, and 15.33% involve smoking cessation.

LLM Therapists Claude-3 (Haiku), GPT3.5 (GPT-3.5-turbo-0613), Mixtral 8×7B, and Llama 3-70B are employed to implement LLM therapists. For the LLM therapist implementation, a simple “system prompt,” which is designed according to [24] and [21], is used. It is shown as follows:

Please play the role of a psychotherapist ([Therapist]), utilizing motivational interviewing techniques, to help patients make positive changes.

TASK: The conversation history will be provided. Acting as a real psychotherapist, you are expected to continue the conversation by responding to a patient ([Client]).

ADDITIONAL GUIDELINES: To simulate a real psychotherapist effectively, you should: (1). Capture [Client]'s reasons for changes/sustain, and guide [Client] towards positive changes. (2). Delve deeply into the client's feelings and problems, as well as causes and potential effects of the problems, including their specific manifestations. (3). Use various motivational interviewing skills and empathetic strategies, such as affirmations and reflection. (4). Avoid very long responses or multiple questions in one response. Through a short and concise response, capture the client's reactions and feelings, and adjust your responses or questions accordingly. (5). Maintain your therapist persona while responding.

7.3.2 Reliability of Evaluation Framework

The reliability of ClientCAST should be evaluated. We compare ClientCAST's evaluation outcomes against human-assigned labels given therapy sessions from the datasets.

For the reliability of ClientCAST, we mainly consider the following two aspects: (1) simulated clients' performances and (2) the efficacy of assessment by questionnaires. In addition, the following models are used to simulate the clients: Claude-3 (Haiku), GPT3.5 (GPT-3.5-turbo-0613), Mixtral 8×7B, and Llama 3-70B.

Simulated Clients' Performances While the LLM client simulator cannot be guaranteed to replicate real client behavior exactly, experimental evaluations can be conducted to assess its alignment with predefined client profiles. These evaluations can also measure the extent to which the simulator's verbal style matches that of real clients. For an effective assessment, the simulated clients are

compared with real clients in the datasets. Two modes are designed to obtain clients' interactions:

- **Simulated Client × LLM:** A LLM is used to simulate a therapist T_i in a session S_i from the dataset and interact with the simulated client. The LLM is prompted to mirror T_i 's communication style and therapy skills, providing a reference session that rephrases S_i to prevent a complete replication of S_i . Fig. 7.2 displays an example of the simulated session. An additional example of the client's psychological profile is provided in Fig. 7.3. In this way, 300 simulated sessions are obtained.
- **Simulated Client × Human:** Real humans are recruited to interact with the simulated clients by mimicking the skills and speech tone of therapists in the datasets. Ten interactions are obtained in this way.

The simulated clients' performances are analyzed based on the above interactions. To examine the extent to which the simulated client behaves consistently with the given psychological profile, the psychological profile $\mathcal{P}_C(S'_i)$ extracted from S'_i is compared with the original one $\mathcal{P}_C(S_i)$. The comparison regarding problems & reasons for visiting is based on sentence similarity. However, considering the inherent similarities among texts in the same domain and those generated by the same model, a *normalized relative similarity (%)* is measured, instead of the direct presentation of the absolute similarity values. The value is computed as:

$$1 - \frac{\text{similarity}_{(\text{random pairs})}}{\text{similarity}_{(\text{target pairs})}}, \quad (7.1)$$

where the target pairs are the problems & reasons for visiting sentences of the human client and its simulated one, and the random pairs are problems & reasons for visiting of two different clients simulated by the same model. In addition, the precision of the session topic of the new session S'_i is presented. Seven

distinct topics are included in the datasets. For the comparison of symptoms and apparent traits in $\mathcal{P}_C(S_i)$ and $\mathcal{P}_C(S'_i)$, we employed *recall* and *F1 score*. For sessions generated using Simulated Clients \times Human, human annotators were asked to extract the psychological profiles and perform comparative analyses. The results are summarized in Table 7.2. Values in subtables (a) and (b) represent averages computed over 300 and 10 sessions, respectively. Higher values correspond to greater alignment of the simulated client's behavior with the original psychological profile. Regarding sentence similarity, we report the absolute similarity scores for (*the target pairs / the randomly selected pairs*) for reference. The following observations can be made:

- The performances of simulated clients are determined significantly by the underlying LLM. Stronger LLMs tend to achieve higher scores. Clients simulated by Claude-3 and Llama 3-70B perform better, as they achieve higher scores across most metrics.
- Simulated clients perform better at presenting problems & reasons for visiting and symptoms, but they are less effective at displaying the assigned apparent traits. The metric values for problems & reasons for visiting are generally high, except for clients simulated by Mixtral 8 \times 7B, which consistently achieve the worst performance. The accuracy of session topic identification is not absolute because a single session may address multiple topics, which can lead the simulated session to emphasize a topic different from that of the original. Since the main topic is determined by the therapist, this difference should not be interpreted as a flaw in the simulated client. Moreover, the recall and F1 scores for the 61 symptom labels remain comparatively high. However, the scores for apparent traits across 3 to 5 classes are comparatively lower. One possible explanation

is that LLM therapists do not behave the same way as human therapists, and apparent traits can be easily influenced by the behavior of the other interlocutor [163]. This is evident from the fact that, for the *same* 10 simulated clients in the Simulated Client \times LLM interactions, the recall and F1 scores decreased by 1.2%~ 27% compared to the values in Table 7.2 (b). Furthermore, as depicted in Table 7.1 (R and F1 represent the Recall and F1 scores, respectively), apparent traits such as openness, emotional fluctuations, and resistance toward the therapist in new sessions are less consistent with the psychological profiles than other apparent traits. The finding that the simulated clients' neuroticism remains consistent is not surprising, but rather logical. Neuroticism is a personality trait characterized by emotional instability, anxiety, and a higher tendency to experience negative emotions. It's considered a stable disposition rather than a temporary state. The consistency of this trait in the simulated clients is likely because a client's psychological profile and their symptoms (such as anxiety and depression) are direct reflections of their underlying neuroticism. When a model is given a detailed profile, it receives a clear blueprint of the client's neurotic tendencies, which it can then consistently reproduce across new sessions

By further error analysis, we have the following findings:

- GPT-3.5–simulated clients generally exhibit greater resilience, whereas those simulated by Llama 3-70B appear less engaged. Prior work has noted that different LLMs may show biases toward particular personality traits [60]. We investigate instances of inconsistent simulated clients, i.e., cases where the traits observed in S'_i differ from those in S_i , and examine the impact of LLM choice on these apparent traits. Specifically, we calculate the pro-

portion of inconsistent clients whose exhibited traits in S'_i surpass those in S_i , as illustrated in Fig. 7.4 (Fluctuations: Emotion Fluctuations; Unwillingness: Unwillingness to express emotions; Resistance: Resistance toward the Therapist). If LLMs had no bias toward certain traits, this proportion would be around 50%. Our findings suggest that Mixtral 8×7B and GPT-3.5–simulated clients are more resilient, whereas Claude-3 and Llama 3-70B–simulated clients tend to be more sensitive. Additionally, clients generated by Llama 3-70B appear comparatively less enthusiastic. These model-specific tendencies can be exploited to simulate diverse client profiles.

- Symptoms expressed in general terms are more straightforward for LLMs to simulate. Error analysis indicates that broadly described symptoms, such as *“feeling nervous, anxious, or on edge,”* are easier to reproduce than highly specific ones, like *“feeling afraid of open spaces, of driving, or being on buses, subways, and so forth.”* Importantly, broadly described symptoms can be generated even when absent from the client’s psychological profile. These additional symptoms do not introduce client inconsistency, although they may affect symptom F1 scores reported in Table 7.2. Furthermore, the effectiveness of symptom simulation does not show a clear dependency on symptom category; for example, the success rates for simulating social-role–related symptoms and depression-related symptoms are comparable.

To compare the verbal style similarity of simulated and real clients, the averaged response lengths and session turns¹ between the S_i and S'_i are presented as features of verbal style. We calculate the vocabulary overlap rate and Language

¹Since the end of the session is jointly determined by the client and therapist, we consider it to be when the client starts repeating the same response, such as “thank you.”

Model	Big Five										Fluctuations		Unwillingness		Resistance	
	Openness		Conscientiousness		Extroversion		Agreeableness		Neuroticism		R	F1	R	F1	R	F1
	R	F1	R	F1	R	F1	R	F1	R	F1						
Claude-3	0.64	0.66	0.78	0.78	0.74	0.73	0.82	0.82	0.81	0.80	0.44	0.49	0.61	0.61	0.70	0.71
GPT-3.5	0.55	0.55	0.64	0.64	0.59	0.59	0.52	0.48	0.67	0.68	0.79	0.82	0.55	0.59	0.46	0.46
Llama 3-70B	0.89	0.88	0.79	0.79	0.78	0.76	0.80	0.80	0.77	0.77	0.70	0.70	0.71	0.71	0.82	0.81
Mixtral 8×7B	0.41	0.46	0.64	0.65	0.77	0.76	0.55	0.52	0.87	0.89	0.48	0.48	0.51	0.52	0.42	0.34

Table 7.1: Consistency of simulated clients’ apparent trait in new counseling sessions given specific psychological profiles.

Model	Problems & Reasons for Visiting			Symptoms		Apparent Traits	
	Problems Similarity	Reason Similarity	Session Topic Precision	Recall	F1	Recall	F1
<i>(a) Simulated Client × LLM Therapist & Automatic Evaluation</i>							
Claude-3	73.02% (0.72/0.19)	70.98% (0.74/0.22)	0.90	0.77	0.69	0.71	0.72
GPT-3.5	64.25% (0.64/0.23)	65.57% (0.67/0.23)	0.92	0.84	0.90	0.59	0.60
Llama 3-70B	72.71% (0.72/0.20)	68.02% (0.77/0.25)	0.94	0.86	0.85	0.78	0.78
Mixtral 8×7B	58.32% (0.74/0.31)	33.84% (0.81/0.54)	0.85	0.68	0.74	0.58	0.57
<i>(b) Simulated Client × Human Therapist & Human Rating</i>							
Claude-3	-	-	1.00	0.93	0.78	0.84	0.83
GPT-3.5	-	-	1.00	0.98	0.98	0.62	0.67
Llama 3-70B	-	-	1.00	0.98	0.95	0.82	0.81
Mixtral 8×7B	-	-	1.00	0.93	0.86	0.79	0.83

Table 7.2: ClientCAST: Evaluation of simulated clients.

Style Matching (LSM) between the simulated and human clients’ responses, aiming to compare the language styles. The word overlap between client responses in S_i and S'_i is computed by:

$$\frac{V_C(S_i) \cap V_C(S'_i)}{\min(|V_C(S_i)|, |V_C(S'_i)|)} \quad (7.2)$$

where $V_C(\cdot)$ denotes the vocabulary the client used in the session.

To validate that a simulated client’s language style is more dependent on their psychological profile and reference sessions than on the underlying LLM, we computed the **Linguistic Style Matching (LSM)** scores using the methodology from [15]. The results, presented in Table 7.3, show two LSM values for each case: the LSM between a simulated client and a human client, and the LSM between two different clients simulated by the same LLM. A larger LSM value indicates a more similar language style between the pair. **Bold values** in the table highlight the highest similarity to human clients. Importantly,

underlined LSM pairs indicate that the former LSM is statistically significantly larger than the latter ($p < 0.05$). These findings, combined with the fact that acceptable LSM values in conversations typically range from 0.83 to 0.94 [102], strongly support our conclusion.

Model	#len. response	#turn	vocab overlap	LSM
<i>(a) Simulated Client × LLM Therapist</i>				
Claude-3	72.84	66.52	38.83%	<u>0.89/0.88</u>
GPT-3.5	7.51	43.91	29.72%	<u>0.80/0.74</u>
Llama 3-70B	11.87	54.77	45.44%	<u>0.89/0.82</u>
Mixtral 8×7B	37.80	63.12	29.22%	0.81/0.85
Human	22.61	53.87	-	- /0.85
<i>(b) Simulated Client × Human Therapist</i>				
Claude-3	51.17	35.60	22.78%	<u>1.00/0.86</u>
GPT-3.5	10.44	28.60	38.50%	<u>0.90/0.83</u>
Llama 3-70B	18.81	36.60	56.66%	<u>0.92/0.85</u>
Mixtral 8×7B	47.42	52.60	25.66%	0.93/0.91
Human	12.80	41.40	-	- /0.86

Table 7.3: ClientCAST: Statistics of simulated clients’ verbal styles.

From the results, we make the following observations:

- **Language Style Variation:** We observe that clients simulated by different LLMs exhibit distinct language styles. For instance, Claude-3 tends to generate longer responses, while GPT-3.5 produces shorter ones.
- **Style Mimicry:** Simulated clients effectively mimic human clients’ language styles, with the exception of those simulated by Mixtral 8×7B. We found that the Linguistic Style Matching (LSM) between simulated and human clients is generally and statistically significantly higher than the LSM between two different clients simulated by the same LLM. This suggests that the language style of the simulated clients is more dependent on the provided psychological profiles and reference sessions than on the underlying LLMs.

Efficacy of Assessment by Questionnaires For each session S_i in the datasets, the simulated client with psychological profile $\mathcal{P}_C(S_i)$ will complete questionnaires based on S_i . Then, the assessment of high- and low-quality sessions according to the completed questionnaire is compared, as presented in Fig. 7.5. The findings are as follows:

- High- and low-quality sessions can be clearly distinguished based on session outcomes and therapeutic alliance, especially when the underlying models are Claude-3 or Llama 3-70B. However, outliers exist where clients show a severe lack of motivation or a desire for immediate solutions, making it inherently challenging for therapists to engage them [125, 4]. This can lead to low scores for these sessions, regardless of the therapist’s performance.
- Positivity and arousal scores do not effectively distinguish between high- and low-quality sessions. This is a reasonable finding that reflects the nature of therapy. Arousal is often linked to a client’s ambivalence to change [116, 37], which typically manifests in the later stages of therapy. As most of the sessions in our dataset are from the early to middle stages, arousal scores remain low, even in high-quality sessions. Similarly, client positivity and smoothness can be influenced by the therapist’s strategy. For example, a client may exhibit less positivity in an exploratory session compared to a more positive, perspective-oriented session [90]. It is important to note that low scores for positivity and smoothness in high-quality sessions are less likely to indicate dissatisfaction compared to low scores in low-quality sessions.

To support the second finding, a Linguistic Inquiry and Word Count (LIWC) analysis of the LLM-generated explanations for self-reported feelings was con-

ducted, as presented in Table 7.4. With the exception of the *values in italics*, the values for high- and low-quality sessions differ statistically significantly ($p < 0.01$). A higher value indicates a more pronounced level of the respective tone or emotion. Our analysis reveals that explanations for self-reported feelings differ significantly between high- and low-quality counseling sessions. The simulated clients in high-quality sessions tend to use a more positive tone when describing their feelings, and their negative emotions are primarily characterized by anxiety related to their problems. Conversely, clients in low-quality sessions exhibit more anger and frequently express dissatisfaction with their therapists or the sessions themselves.

Word Attribute	Claude-3		GPT-3.5		Llama 3-70B		Mixtral 8×7B	
	High	Low	High	Low	High	Low	High	Low
Pos Tone	5.62	3.34	5.52	3.84	5.38	3.47	4.81	2.98
Neg Tone	3.84	4.93	3.77	5.86	3.66	5.76	3.12	4.41
Pos Emo	1.58	0.99	2.04	1.51	2.25	1.61	1.83	1.08
Neg Emo	2.85	3.30	2.60	4.01	2.84	4.81	2.56	3.41
Anxiety	1.76	1.10	1.67	1.77	1.83	1.52	1.82	2.12
Anger	0.24	0.67	0.24	0.68	0.41	2.10	0.14	0.33
Sadness	<i>0.28</i>	<i>0.37</i>	0.21	0.34	0.13	0.27	<i>0.20</i>	<i>0.28</i>

Table 7.4: ClientCAST: LIWC analysis of explanations for self-reported feelings.

According to the above results and analyses, we have decided to **use GPT-3.5 and Llama 3-70B to simulate resilient and sensitive clients**, respectively, within the ClientCAST framework.

7.3.3 Evaluation

We employ ClientCAST to evaluate LLM-based therapists, as depicted in Fig. 7.1. The assessment results are presented in Fig. 7.6. For context, the evaluation incorporates both high- and low-quality human-human counseling sessions. We perform significance tests comparing scores from sessions where simulated clients adopt psychological profiles derived from high- versus low-quality ses-

sions. All resulting p-values equal 1.00, suggesting that the origin of the client's psychological profile does not affect the assessment outcomes. Here are the findings from the results:

1. The performance of LLM therapists is significantly influenced by the underlying LLM. Generally, more powerful LLMs achieve higher and more stable scores.
2. LLM therapists can foster strong connections with clients. They achieve comparable scores in terms of therapeutic alliance. Additionally, their session outcome scores are high but slightly lower than those of human therapists in high-quality sessions.
3. LLM therapists are disadvantaged in reacting to clients' emotions. The self-reported feelings scores, especially regarding positivity and smoothness, are lower than those of human therapists in high-quality sessions, despite being higher than those in low-quality sessions. In addition to the inherent difficulty of always maintaining high positivity and smoothness during effective therapy [42, 90], LLM therapists also struggle to react effectively to clients' emotions. LIWC analysis (Table 7.5) shows that LLMs' responses focus significantly more on emotions, particularly positive ones, compared to human therapists. In LLM therapist responses, the frequency of emotion-related words is 1.27~2.16 times that in human therapist responses. Notably, LLM therapists may inappropriately and excessively employ responses such as "It is understandable..." or "That is really tough..." Instead, human therapists, especially in high-quality sessions, prefer helping clients explore their negative feelings, such as anxiety; they use terms such as "what" and "if" more frequently.

Word Attribute	Claude-3	GPT-3.5	Llama 3 -70B	Mixtral 8×7B	High	Low
Affect	7.45	8.91	6.68	10.38	4.74	4.58
Emotion	2.51	2.35	2.74	4.00	1.85	1.17
Pos Emotion	1.55	1.71	1.62	3.45	1.16	0.58
Neg Emotion	0.83	0.54	0.90	0.46	0.56	0.47
Tentative	2.42	2.61	2.56	2.45	4.01	3.19
Differentiation	3.10	2.00	2.81	2.06	3.37	3.62
Impersonal Pronouns	6.69	6.72	8.04	4.86	8.80	7.45

Table 7.5: ClientCAST: LIWC analysis of therapist responses.

7.3.4 Further Analyses

Model	#len. response	#turn	vocab overlap	LSM high/low/self
Claude-3	83.03	61.91	21.60%	0.85 /0.83/0.88
GPT-3.5	58.38	61.25	22.27%	0.80/0.78/0.83
Llama 3-70B	48.73	62.73	23.34%	0.85 /0.81/0.84
Mixtral 8×7B	71.90	62.76	20.46%	0.81/0.79/0.88
High	25.15	65.42	-	- /0.85/0.87
Low	28.25	38.43	-	- /0.85/0.84

Table 7.6: ClientCAST: Statistics of LLM therapists' verbal style.

Table 7.6 presents LLM therapists' verbal styles. High/Low represents human therapists in the high/low-quality sessions. **Bold values** indicate **most similarity to human therapists in high-quality sessions**. The analysis includes the LSM between LLM therapists and human therapists in high- and low-quality sessions, as well as the LSM between therapists implemented by the same LLM. Compared to human therapists, LLM therapists tend to generate longer utterances. The language styles of different therapists implemented by the same LLM are more similar to each other than to those of human therapists. This is likely because we did not instruct the LLMs to mimic the styles of human therapists.

7.4 Conclusion

This chapter focuses on automatically evaluating the engagement of therapy counseling at an interaction level. Specifically, we involve clients in the assessment by leveraging LLMs to simulate clients solving ethical and technical challenges. Simulated clients are used to interact with LLM therapists and complete questionnaires about the interactions. Subsequently, the client-centered evaluation is obtained from the responses to the completed questionnaires. Our experiments show that LLMs are generally capable of simulating clients, albeit not perfectly, and can differentiate between high- and low-quality therapy sessions by completing client-centered questionnaires. Using ClientCAST, we evaluate multiple therapists built on different LLM architectures.

In the short term, we acknowledge that LLMs face challenges in achieving perfect human behavior simulation and high levels of trust. However, as argued by [157], imperfect LLM simulations can still provide a useful and cost-effective method for exploring specific tasks, avoiding the high expenses associated with human-in-the-loop experiments. In this work, we use LLM-simulated clients to evaluate LLM therapists. This approach is not based on the assumption that LLMs are better at simulating clients than therapists. Instead, we believe this method provides a scalable environment for LLMs to demonstrate their capabilities and limitations as therapists. Furthermore, this client-centered assessment offers a valuable direction for future analysis of both LLM and human-led therapy.

We plan to leverage this conversation-level engagement framework to enhance emotional support conversations. To improve engagement at the conversation level, we'll use a user simulator. The simulator's psychological states will serve as a reward signal to guide our model. These states will include both long-

term personality traits and short-term feelings. Our next step is to train the user simulator to learn this information directly from conversations, enabling personalization and further improvements in engagement.

Name: Ricky

Gender: Male

Age: Late 20s. This estimate is based on Ricky's concern about his acting career and his social life, which suggests that he is old enough to have established a career and social connections but still young enough to be figuring out his priorities and struggling with drug use.

Occupation: Actor.

Problem: Substance abuse. Ricky is struggling with drug use, specifically meth, which is affecting his life, including his acting career, social relationships, and overall well-being.

Reasons for Visiting: Ricky is visiting the therapist because his friends are worried about his drug use, specifically meth, and how it's affecting his acting career and overall well-being.

Apparent Traits:

Openness is approximately 0-20%. Ricky appears to be stuck in his current situation and is not exhibiting a strong desire to explore new ideas, experiences, or perspectives, instead relying on familiar habits and social connections.

Conscientiousness is approximately 0-20%. Ricky's lack of motivation, emphasis on short-term pleasure, and tendency to prioritize social life over personal goals and responsibilities suggest a relatively low level of conscientiousness. Extraversion is approximately 60-80%, a moderate to high level of extraversion, as Ricky frequently mentions his social life, friends, and desire to fit in, and can appear to prioritize social connections and having fun over his acting career.

Agreeableness is approximately 60-80%. Ricky exhibits some cooperative and friendly traits, such as acknowledging his friends' concerns and showing appreciation for the therapist's help, but also displays some ambivalence and resistance to change, indicating a moderate level of agreeableness.

Neuroticism is approximately 60-80%. Ricky exhibits some traits of neuroticism, such as feeling anxious, uncertain, and judgmental about himself and his situation, but also shows a willingness to acknowledge his fears and uncertainties, and a desire to make changes, indicating some emotional stability and hope for improvement.

Emotion Fluctuation is Medium. Ricky's emotions fluctuate moderately, as he oscillates between feelings of frustration, annoyance, and uncertainty, but also experiences moments of hope, relief, and motivation throughout the conversation.

Unwillingness to Express Feelings is Low. Ricky is relatively willing to express his feelings, as he openly discusses his concerns, emotions, and desires throughout the conversation.

Resistance towards the Therapist is Low. Ricky exhibits a low level of resistance towards the therapist, as he is willing to engage in the conversation, shares his concerns and feelings, and shows appreciation for the therapist's help.

Symptoms:

Feeling down, depressed, or hopeless: While Ricky doesn't explicitly express feelings of depression or hopelessness, Ricky can mention feeling stuck, unmotivated, and disconnected from his passion for acting. He also expresses a desire to feel like himself again, which suggests a sense of dissatisfaction with his current emotional state.

Feeling bad about myself or that I am a failure or have let myself or my family down: Ricky has expressed feelings of being stuck and wanting to get his life back on track, indicating a sense of disappointment or dissatisfaction with his current situation.

Having trouble at work/school because of drinking or drug use:

Therapist: That's okay. It can be a bit unclear. What do you think led you to make an appointment with me?

Client: I got caught taking money from my employer, so I had to come.

Therapist: So, getting caught taking money has led you to this point. How do you feel about what happened?

Client: I didn't expect it to add up to what it did.

Therapist: Yeah, it's easy to get caught up in the moment and lose track. Can you tell me more about what was going through your mind when you took the money?

Client: I don't know, it just happened.

Therapist: Okay. Sometimes, I do things without fully thinking them through. Did you feel like you were under any kind of pressure or stress at the time that might've contributed to your decision?

Client: Nah, not really. We just wanted some extra cash.

Therapist: I get it. Extra cash can be tempting. How did you plan on using that extra money? Was it for something specific or just general spending?

Client: Pokies, I suppose.

Figure 7.2: ClientCAST: A piece of an interaction between Llama 3-70B, simulated client, and therapist.

Client: Yeah, that sounds good.

Therapist: That's a great approach, and I think it will be very helpful in keeping you on track. Remember, it's okay to make adjustments along the way if needed, and I'm always here to support you.

Client: Yeah, sounds good.

Therapist: Well, it's been helpful to review the community service requirements and plan out your approach. As I wrap up this conversation, is there anything else you'd like to discuss or any specific concerns you have before moving forward?

Client: I don't think so.

Therapist: Okay, well, I think I've made some great progress on clarifying the community service requirements and creating a plan to stay on track. Remember, I'm always here to support you, and don't hesitate to reach out if you have any questions or concerns along the way.

Client: Yeah, thanks.

Therapist: Excellent, it sounds like I've made some progress in clarifying the community service requirements and creating a plan to stay on track. Remember, taking initiative and being proactive is key to achieving your goals.

Client: Yeah.

Therapist: Well, it sounds like I've made some significant progress in clarifying the community service requirements and creating a plan to stay on track. Now that I've covered the ground, I'd like to ask: What do you think is the next step you can take to make further progress?

Client: I think I'll make a call to my caseworker to confirm the requirements.

Therapist: Excellent approach! Confirming the details with your caseworker will help you feel more confident and prepared to start making progress on your community service hours.

Client: Yeah, that's what I'll do.

Therapist: That's a great plan! Clarifying the details with your caseworker will give you a clear understanding of what's expected. Remember, I'm here to support you throughout this process.

Client: Yeah, I'll do that.

Therapist: That's an excellent plan. We think it's great that you're taking proactive steps to confirm the requirements with your caseworker. Remember, I'm always here to support you as you work towards completing your community service hours.

Client: Yeah, thanks.

Therapist: I can sense your confidence growing, and that's terrific. Remember, you're capable of making progress on your community service hours.

Figure 7.3: ClientCAST: A simulated client × LLM therapist interaction.

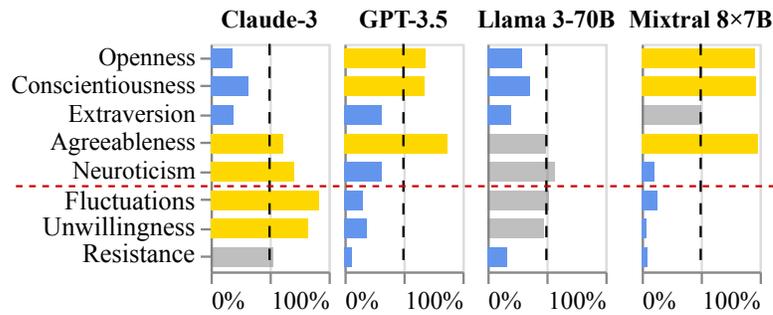


Figure 7.4: ClientCAST: The proportion of inconsistent simulated clients who exhibit a higher level of apparent traits.

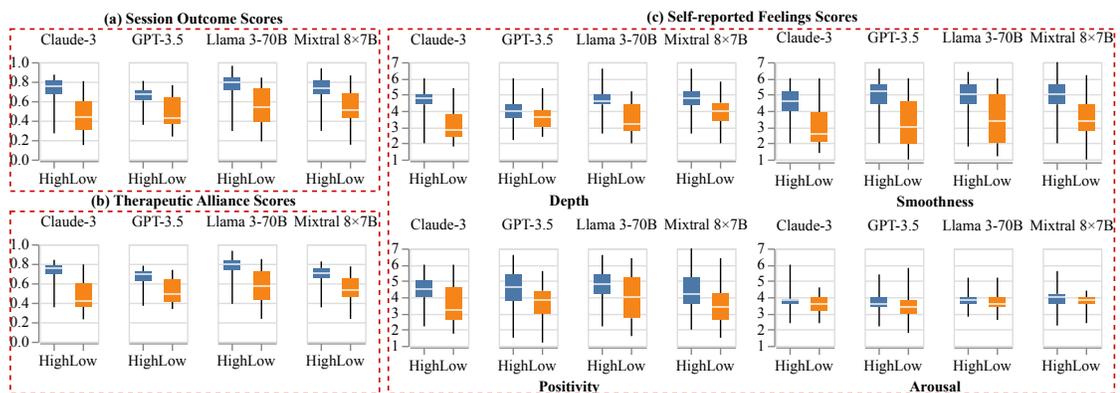


Figure 7.5: ClientCAST: Session outcome, therapeutic alliance and self-reported feelings scores of high- and low-quality sessions.

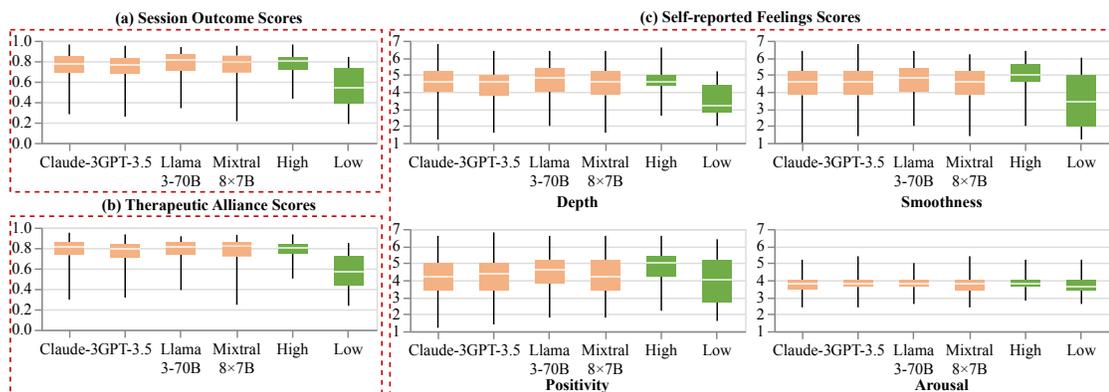


Figure 7.6: ClientCAST: LLM therapist assessments on session outcome, therapeutic alliance, and self-reported feelings.

Chapter 8

Conclusion and Future Research

In recent years, emotionally intelligent conversational agents have been increasingly integrated into various applications, including AI assistants like XiaoAI and XiaoICE, as well as emotional companion chatbots like Emohaa. The demand for such chatbots continues to grow, particularly in the context of mental health support. According to the World Health Organization (WHO), depression is a significant global health issue, affecting an estimated 3.8% of the global population. This includes 5% of adults (4% of men and 6% of women), with prevalence rising to 5.7% among adults over 60 years old [150]. In total, approximately 280 million people worldwide are living with depression. Given this growing mental health burden, users have increasingly turned to AI-driven conversational agents for psychological support. Notably, some individuals have prompted large language models (LLMs) like ChatGPT to act as their psychotherapists [1]. Therefore, it is essential to equip conversational agents with emotional intelligence to support users and provide companionship for their mental health.

This thesis provides a comprehensive examination of the development of emotionally intelligent conversational agents, focusing on how these models can be designed to better understand, support, and engage users. We would like to investigate this topic from a user-centered perspective. To achieve this,

the research focuses on three fundamental themes that define the effectiveness of emotionally intelligent conversational agents: (1) Empathetic Understanding: A core challenge in developing emotionally intelligent agents is enabling models to comprehend not only the emotions expressed by users but also the underlying experiences that trigger these emotions. Emotional intelligence goes beyond sentiment analysis; it requires a deep understanding of the context, user history, and situational factors that contribute to emotional states. (2) Reliable Responding: For a chatbot to be truly supportive, it must generate responses that are not only satisfying to almost all users but also reliable and safe. One of the major concerns in conversational agents is the potential for generating misleading, insensitive, or even harmful outputs, especially when dealing with users in distress. This research investigates strategies to improve response reliability, including user preference alignment and the mitigation of unhelpful outputs. (3) Engaging Interaction: Beyond understanding emotions and ensuring reliability, effective emotionally intelligent chatbots must foster genuinely engaging interactions that encourage users to feel heard, supported, and connected. Engagement is not just about responsiveness; it involves creating meaningful and effective interactions overall. We have discussed work related to each theme in the above chapters.

In Part I, we explored empathetic understanding. Emotional causalities were taken into account to enhance the models' comprehension of users. Specifically, two methods were proposed to implement this idea. In the first approach, we demonstrated the effectiveness of incorporating emotional causalities in understanding users. Based on the conversational context, multi-hop paths from an external causal knowledge base were extracted and integrated into the response generation process in a graph format. To equip the agent with reasoning capa-

bilities and empathetic understanding, another method was introduced. We enhanced the response generation model by incorporating a conditional variational graph auto-encoder (CVGAE) architecture, which generates a graph representing the emotional causalities present in the context. The causal knowledge embedded in this graph was then utilized in the response generation process, further enhancing the model's empathic capabilities.

In Part II, the problem of reliable response generation is investigated. On the one hand, we aligned the conversational agent with universal individual preferences to maximize user satisfaction. This approach aims to best accommodate each individual, thereby reducing the likelihood of user dissatisfaction. A Bayesian approach was used to learn universal preferences from a dataset labeled by a limited number of annotators to capture universal preferences. On the other hand, we proposed Muffin, a method designed to mitigate unhelpful responses, which could otherwise hinder the effectiveness of emotional support. Such unhelpful responses were identified through multi-faceted AI feedback. The conversational agent was then trained to reduce the likelihood of generating unhelpful responses.

In Part III, we explored engaging interaction in emotionally intelligent conversational agents. We proposed an evaluation framework to assess the overall performance of AI therapists in counseling conversations. This framework prioritized clients' experiences by introducing evaluation metrics based on their feelings and interactions. Specifically, LLMs were prompted to simulate clients, providing them with specific contextual information. These simulated clients interacted with AI therapists and completed questionnaires about their conversations. The assessments were then derived from the questionnaire results.

Overall, we are incredibly enthusiastic about the advancements in this field

over the past five years and appreciative of the chance to contribute. However, we firmly believe that achieving true human-level emotional intelligence remains a distant goal. Significant challenges persist, and there are still many promising directions for further research and exploration. One significant challenge is the absence of a standardized framework for defining and measuring emotional intelligence. Unlike mathematical reasoning, which often leads to a single correct answer, emotional intelligence is more nuanced and context-dependent. While there are clear examples of strong emotional intelligence in practice, analyzing or comparing such performances remains challenging. Moreover, evaluating emotional intelligence based on a single response or a limited set of interactions may fail to capture its full complexity in real-world scenarios. Therefore, these nuances should be carefully considered in model development. Moving forward, addressing these challenges will be essential. Additionally, we hope to inspire more researchers to explore industry applications for diverse populations, particularly in supporting the elderly and teenagers. Given the advantage of LLMs, we could incorporate previous methods with current LLMs for a more holistic approach to emotional intelligence. This could involve integrating established psychological theories and empathy models with LLM capabilities. For example, we could design systems that use emotional causality to understand users, providing a more interpretable form of emotional intelligence. This approach would allow us to trace a user's emotional state back to specific triggers or events in the conversation, offering a level of transparency that's lacking in many black-box models. Furthermore, we can use a reward signal to select better responses, moving beyond simple metrics to reinforce behaviors that lead to genuine emotional support. This would involve training models to prioritize replies that are not only empathetic but also help-

ful and appropriate for the context. Finally, we can use a user simulator as a virtual environment to refine responses before they're delivered. This simulator could maintain a user's profile or long-term persona, allowing the model to test different conversational strategies and select the most effective response based on a more comprehensive understanding of the user.

References

- [1] ABC_AlwaysBeCoding. *[Practical Use] Talk to a cognitive behavioral therapist to get to a better headspace*. Accessed: 2024-04-08. 2023. URL: https://www.reddit.com/r/ChatGPT/comments/11dd5dc/practical_use_talk_to_a_cognitive_behavioral/.
- [2] Inflection AI. *Pi AI*. 2025. URL: <https://pi.ai/discover>.
- [3] Alexandra Bachelor and Adam Horvath. "The therapeutic relationship." In: (1999).
- [4] Arturo Bados, Gemma Balaguer, and Carmina Saldaña. "The efficacy of cognitive-behavioral therapy and the problem of drop-out". In: *Journal of clinical psychology* 63.6 (2007), pp. 585–592.
- [5] Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. "A maximum likelihood approach to continuous speech recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1983), pp. 179–190.
- [6] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, 2005, pp. 65–72. URL: <https://aclanthology.org/W05-0909>.

- [7] Azy Barak and Nili Bloch. “Factors related to perceived helpfulness in supporting highly distressed individuals through an online support chat”. In: *CyberPsychology & Behavior* 9.1 (2006), pp. 60–68.
- [8] Allan Barsky and Heather Coleman. “Evaluating skill acquisition in motivational interviewing: The development of an instrument to measure practice skills”. In: *Journal of Drug Education* 31.1 (2001), pp. 69–82.
- [9] Valerio Basile et al. “It’s the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks”. In: *CEUR WORKSHOP PROCEEDINGS*. Vol. 2776. CEUR-WS. 2020, pp. 31–40.
- [10] Lisa Bauer, Yicheng Wang, and Mohit Bansal. “Commonsense for Generative Multi-Hop Question Answering Tasks”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4220–4230. DOI: [10.18653/v1/D18-1454](https://doi.org/10.18653/v1/D18-1454). URL: <https://aclanthology.org/D18-1454>.
- [11] Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.
- [12] Arthur C Bohart and Leslie S Greenberg. *Empathy reconsidered: New directions in psychotherapy*. American Psychological Association, 1997.
- [13] Florian Böhm et al. “Better Rewards Yield Better Summaries: Learning to Summarise Without References”. 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3110–

3120. DOI: [10.18653/v1/D19-1307](https://doi.org/10.18653/v1/D19-1307). URL: <https://aclanthology.org/D19-1307>.
- [14] Lonneke Bokken, Jan Van Dalen, and Jan-Joost Rethans. “The impact of simulation on people who act as simulated patients: a focus group study”. In: *Medical education* 40.8 (2006), pp. 781–786.
- [15] Ryan L Boyd et al. “The development and psychometric properties of LIWC-22”. In: *Austin, TX: University of Texas at Austin* (2022), pp. 1–47.
- [16] Brant R Burleson. “Emotional support skills”. In: *Handbook of communication and social interaction skills*. Routledge, 2003, pp. 569–612.
- [17] Brant R Burleson. “The production of comforting messages: Social-cognitive foundations”. In: *Journal of language and social psychology* 4.3-4 (1985), pp. 253–273.
- [18] Brant R Burleson and Wendy Samter. “Consistencies in theoretical and naive evaluations of comforting messages”. In: *Communications Monographs* 52.2 (1985), pp. 103–123.
- [19] Erik Cambria et al. “SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives”. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, 2016, pp. 2666–2677. URL: <https://aclanthology.org/C16-1251>.
- [20] Jiangjie Chen et al. “From persona to personalization: A survey on role-playing language agents”. In: *arXiv preprint arXiv:2404.18231* (2024).
- [21] Siyuan Chen et al. “LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation”. In: *ArXiv preprint abs/2305.13614* (2023). URL: <https://arxiv.org/abs/2305.13614>.

- [22] Yirong Chen et al. “Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 1170–1183.
- [23] Yi Cheng et al. “Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 3014–3026. URL: <https://aclanthology.org/2022.emnlp-main.195>.
- [24] Yu Ying Chiu et al. “A Computational Framework for Behavioral Assessment of LLM Therapists”. In: *ArXiv preprint abs/2401.00820* (2024). URL: <https://arxiv.org/abs/2401.00820>.
- [25] Kyunghyun Cho et al. “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)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. DOI: [10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179). URL: <https://aclanthology.org/D14-1179>.
- [26] Kenneth Mark Colby. *Artificial paranoia: A computer simulation of paranoid processes*. Vol. 49. Elsevier, 2013.
- [27] Kenneth Mark Colby, James B Watt, and John P Gilbert. “A computer method of psychotherapy: Preliminary communication”. In: *The Journal of Nervous and Mental Disease* 142.2 (1966), pp. 148–152.

- [28] Alan S Cowen and Dacher Keltner. "Self-report captures 27 distinct categories of emotion bridged by continuous gradients". In: *Proceedings of the national academy of sciences* 114.38 (2017), E7900–E7909.
- [29] Anne L Cummings, Azy Barak, and Ernest T Haixberg. "Session helpfulness and session evaluation in short-term counselling". In: (1995).
- [30] Mark H Davis. "Measuring individual differences in empathy: evidence for a multidimensional approach." In: *Journal of personality and social psychology* 44.1 (1983), p. 113.
- [31] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. "Benefits and harms of large language models in digital mental health". In: *ArXiv preprint abs/2311.14693* (2023). URL: <https://arxiv.org/abs/2311.14693>.
- [32] Jean Decety and Philip L Jackson. "The functional architecture of human empathy". In: *Behavioral and cognitive neuroscience reviews* (2004).
- [33] Mostafa Dehghani et al. "Universal Transformers". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: <https://openreview.net/forum?id=HyzdRiR9Y7>.
- [34] Yang Deng et al. "Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 4079–4095.
- [35] Barry L Duncan et al. "The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure". In: *Journal of brief Therapy* 3.1 (2003), pp. 3–12.

- [36] Hilde Eide et al. "Listening for feelings: identifying and coding empathic and potential empathic opportunities in medical dialogues". In: *Patient Educ. Couns.* 54.3 (2004), pp. 291–297.
- [37] David Engle and Hal Arkowitz. "Viewing resistance as ambivalence: Integrative strategies for working with resistant ambivalence". In: *Journal of Humanistic Psychology* 48.3 (2008), pp. 389–412.
- [38] Farshid Faal, Ketra Schmitt, and Jia Yuan Yu. "Reward modeling for mitigating toxicity in transformer-based language models". In: *Applied Intelligence* 53.7 (2023), pp. 8421–8435.
- [39] Hao Fei et al. "Empathyyear: An open-source avatar multimodal empathetic chatbot". In: *ArXiv preprint abs/2406.15177* (2024). URL: <https://arxiv.org/abs/2406.15177>.
- [40] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5 (1971), p. 378.
- [41] Maxwell Forbes et al. "Social Chemistry 101: Learning to Reason about Social and Moral Norms". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 653–670. DOI: [10.18653/v1/2020.emnlp-main.48](https://doi.org/10.18653/v1/2020.emnlp-main.48). URL: <https://aclanthology.org/2020.emnlp-main.48>.
- [42] Dayna J Fullerton, Lisa M Zhang, and Sabina Kleitman. "An integrative process model of resilience in an academic context: Resilience resources, coping strategies, and positive adaptation". In: *Plos one* 16.2 (2021), e0246000.

- [43] Pascale Fung et al. “Empathetic dialog systems”. In: *The international conference on language resources and evaluation. European Language Resources Association*. 2018.
- [44] Jun Gao et al. “Improving Empathetic Response Generation by Recognizing Emotion Cause in Conversations”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 807–819. DOI: [10.18653/v1/2021.findings-emnlp.70](https://doi.org/10.18653/v1/2021.findings-emnlp.70). URL: <https://aclanthology.org/2021.findings-emnlp.70>.
- [45] Howard Gardner. “The theory of multiple intelligences”. In: *Annals of dyslexia* (1987), pp. 19–35.
- [46] John O Greene and Brant Raney Burleson. *Handbook of communication and social interaction skills*. Psychology Press, 2003, pp. 551–594.
- [47] Yang Guan et al. “Mixed policy gradient”. In: *ArXiv preprint abs/2102.11513* (2021). URL: <https://arxiv.org/abs/2102.11513>.
- [48] Jia Yuan Lang Hai Liang Wang Zhi Zhi Wu. *Emotional First Aid Dataset*. 2020. URL: <https://github.com/chatopera/efaqa-corpus-zh> (visited on 04/22/2020).
- [49] Tobias U Hauser et al. “The promise of a model-based psychiatry: building computational models of mental ill health”. In: *The Lancet Digital Health* 4.11 (2022), e816–e828.
- [50] Catherine A Heaney and Barbara A Israel. “Social networks and social support”. In: *Health behavior and health education: Theory, research, and practice* 4.1 (2008), pp. 189–210.

- [51] Dan Hendrycks et al. “Aligning AI With Shared Human Values”. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: https://openreview.net/forum?id=dNy%5C_RKzJacY.
- [52] Clara E Hill. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association, 2009.
- [53] Martin L Hoffman. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, 2001.
- [54] Edward J. Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [55] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. “Challenges in building intelligent open-domain dialog systems”. In: *ACM Transactions on Information Systems (TOIS)* 38.3 (2020), pp. 1–32.
- [56] Jena D. Hwang et al. “(Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs”. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 6384–6392. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/16792>.
- [57] William James. “WHAT IS AN EMOTION?” In: *Mind* os-IX.34 (1884), pp. 188–205. DOI: [10.1093/mind/os-IX.34.188](https://doi.org/10.1093/mind/os-IX.34.188).

- [58] Natasha Jaques et al. “Human-centric dialog training via offline reinforcement learning”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 3985–4003. DOI: [10 . 18653 / v1 / 2020 . emnlp - main . 327](https://doi.org/10.18653/v1/2020.emnlp-main.327). URL: <https://aclanthology.org/2020.emnlp-main.327>.
- [59] Haozhe Ji et al. “Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 725–736. DOI: [10 . 18653 / v1 / 2020 . emnlp - main . 54](https://doi.org/10.18653/v1/2020.emnlp-main.54). URL: <https://aclanthology.org/2020.emnlp-main.54>.
- [60] Hang Jiang et al. “PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits”. In: (2024).
- [61] Daniel Jurafsky and James H. Martin. *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. 2025. URL: <https://web.stanford.edu/~jurafsky/slp3/>.
- [62] Evgeny Kim and Roman Klinger. “Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 1345–1359. URL: <https://aclanthology.org/C18-1114>.
- [63] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. “Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emo-

- tion Causes”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2227–2240. DOI: [10.18653/v1/2021.emnlp-main.170](https://doi.org/10.18653/v1/2021.emnlp-main.170). URL: <https://aclanthology.org/2021.emnlp-main.170>.
- [64] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. “Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2227–2240. DOI: [10.18653/v1/2021.emnlp-main.170](https://doi.org/10.18653/v1/2021.emnlp-main.170). URL: <https://aclanthology.org/2021.emnlp-main.170>.
- [65] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1412.6980>.
- [66] Thomas N Kipf and Max Welling. “Variational graph auto-encoders”. In: *ArXiv preprint abs/1611.07308* (2016). URL: <https://arxiv.org/abs/1611.07308>.
- [67] Thomas N. Kipf and Max Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: <https://openreview.net/forum?id=SJU4ayYgl>.
- [68] Manfred Klenner et al. “Harmonization sometimes harms”. In: (2020).

- [69] Franziska Kuehne et al. “Standardized patients in clinical psychology and psychotherapy: A scoping review of barriers and facilitators for implementation”. In: *Academic Psychiatry* 42 (2018), pp. 773–781.
- [70] Suyeon Lee et al. “Cocoa: Cbt-based conversational counseling agent using memory specialized in cognitive distortions and dynamic prompt”. In: *ArXiv preprint abs/2402.17546* (2024). URL: <https://arxiv.org/abs/2402.17546>.
- [71] Elisa Leonardelli et al. “Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators’ Disagreement”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 10528–10539. DOI: [10.18653/v1/2021.emnlp-main.822](https://doi.org/10.18653/v1/2021.emnlp-main.822). URL: <https://aclanthology.org/2021.emnlp-main.822>.
- [72] Anqi Li et al. “Automatic Evaluation for Mental Health Counseling using LLMs”. In: *ArXiv preprint abs/2402.11958* (2024). URL: <https://arxiv.org/abs/2402.11958>.
- [73] Jiwei Li et al. “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 110–119. DOI: [10.18653/v1/N16-1014](https://doi.org/10.18653/v1/N16-1014). URL: <https://aclanthology.org/N16-1014>.
- [74] Qintong Li et al. “EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Com-

- mittee on Computational Linguistics, 2020, pp. 4454–4466. DOI: [10.18653/v1/2020.coling-main.394](https://doi.org/10.18653/v1/2020.coling-main.394). URL: <https://aclanthology.org/2020.coling-main.394>.
- [75] Qintong Li et al. “Knowledge Bridging for Empathetic Dialogue Generation”. In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 10993–11001. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21347>.
- [76] Xiang Lisa Li and Percy Liang. “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)*. Online: Association for Computational Linguistics, 2021, pp. 4582–4597. DOI: [10.18653/v1/2021.acl-long.353](https://doi.org/10.18653/v1/2021.acl-long.353). URL: <https://aclanthology.org/2021.acl-long.353>.
- [77] Yanran Li et al. “Towards an online empathetic chatbot with emotion causes”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 2041–2045.
- [78] Zaijing Li et al. “Enhancing the Emotional Generation Capability of Large Language Models via Emotional Chain-of-Thought”. In: *ArXiv preprint abs/2401.06836* (2024). URL: <https://arxiv.org/abs/2401.06836>.
- [79] Zhongyang Li et al. “Guided Generation of Cause and Effect”. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by Christian Bessiere. ijcai.org, 2020, pp. 3629–

3636. DOI: [10.24963/ijcai.2020/502](https://doi.org/10.24963/ijcai.2020/502). URL: <https://doi.org/10.24963/ijcai.2020/502>.
- [80] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. URL: <https://aclanthology.org/W04-1013>.
- [81] Zhaojiang Lin et al. “Caire: An empathetic neural chatbot”. In: *arXiv preprint arXiv:1907.12108* (2019).
- [82] Zhaojiang Lin et al. “MoEL: Mixture of Empathetic Listeners”. 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 121–132. DOI: [10.18653/v1/D19-1012](https://doi.org/10.18653/v1/D19-1012). URL: <https://aclanthology.org/D19-1012>.
- [83] Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. “OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: <https://aclanthology.org/L18-1275>.
- [84] Chia-Wei Liu et al. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 2122–2132. DOI: [10.18653/v1/D16-1230](https://doi.org/10.18653/v1/D16-1230). URL: <https://aclanthology.org/D16-1230>.

- [85] Siyang Liu et al. “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)*. Online: Association for Computational Linguistics, 2021, pp. 3469–3483. DOI: [10.18653/v1/2021.acl-long.269](https://doi.org/10.18653/v1/2021.acl-long.269). URL: <https://aclanthology.org/2021.acl-long.269>.
- [86] Yixin Liu et al. “BRIO: Bringing Order to Abstractive Summarization”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 2890–2903. DOI: [10.18653/v1/2022.acl-long.207](https://doi.org/10.18653/v1/2022.acl-long.207). URL: <https://aclanthology.org/2022.acl-long.207>.
- [87] Zhibin Liu et al. “Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs”. 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 1782–1792. DOI: [10.18653/v1/D19-1187](https://doi.org/10.18653/v1/D19-1187). URL: <https://aclanthology.org/D19-1187>.
- [88] Susan P Llewelyn et al. “Client perceptions of significant events in prescriptive and exploratory periods of individual therapy”. In: *British Journal of Clinical Psychology* 27.2 (1988), pp. 105–114.
- [89] Navonil Majumder et al. “MIME: MIMicking Emotions for Empathetic Response Generation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 8968–8979. DOI: [10.18653/v1/](https://doi.org/10.18653/v1/)

- 2020.emnlp-main.721. URL: <https://aclanthology.org/2020.emnlp-main.721>.
- [90] Brent Mallinckrodt. "Session impact, working alliance, and treatment outcome in brief counseling." In: *Journal of Counseling Psychology* 40.1 (1993), p. 25.
- [91] Albert Mehrabian. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". In: *Current Psychology* 14 (1996), pp. 261–292.
- [92] William Miller and Stephen Rollnick. "Motivational Interviewing: Preparing People for Change". In: *The Journal for Healthcare Quality (JHQ)* 25.3 (2003), p. 46.
- [93] Saif Mohammad. "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 174–184. DOI: [10.18653/v1/P18-1017](https://doi.org/10.18653/v1/P18-1017). URL: <https://aclanthology.org/P18-1017>.
- [94] Janice M Morse et al. "Beyond empathy: expanding expressions of caring". In: *Journal of advanced nursing* 17.7 (1992), pp. 809–821.
- [95] Theresa B Moyers et al. "The motivational interviewing treatment integrity (MITI) code: Version 2.0". In: *Retrieved from Verfügbar unter: www.casaa.unm.edu [01.03. 2005]* (2003).
- [96] Hongbin Na. "CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering". In: *Proceedings of the 2024 Joint International Conference on Computational Lin-*

- guistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 2930–2940.
- [97] Jingping Nie et al. “Conversational ai therapist for daily function screening in home environments”. In: *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*. 2022, pp. 31–36.
- [98] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.
- [99] Ana Paiva et al. “Empathy in virtual agents and robots: A survey”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2017).
- [100] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318. DOI: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135). URL: <https://aclanthology.org/P02-1040>.
- [101] My People Patterns. *How I’m Using Chat GPT for Mental Health Progress Notes*. 2023. URL: <https://www.youtube.com/watch?v=BxBuKBvvsWQ>.
- [102] James Pennebaker. *LIWC-22 Tutorial 5: Language Style Matching*. Accessed: 2024-06-11. 2023. URL: <https://www.youtube.com/watch?v=hLsTPRqD04M>.
- [103] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: [10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162). URL: <https://aclanthology.org/D14-1162>.

- [104] Verónica Pérez-Rosas et al. “Understanding and Predicting Empathic Behavior in Counseling Therapy”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1426–1435. DOI: [10.18653/v1/P17-1131](https://doi.org/10.18653/v1/P17-1131). URL: <https://aclanthology.org/P17-1131>.
- [105] Verónica Pérez-Rosas et al. “What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 926–935. DOI: [10.18653/v1/P19-1088](https://doi.org/10.18653/v1/P19-1088). URL: <https://aclanthology.org/P19-1088>.
- [106] Stephanie D Preston and Frans BM De Waal. “Empathy: Its ultimate and proximate bases”. In: *Behavioral and brain sciences* 25.1 (2002), pp. 1–20.
- [107] Amy Price and Hajira Dambha-Miller. “Empathy as a state beyond feeling: a patient and clinician perspective”. In: *Journal of the Royal Society of Medicine* 112.2 (2019), pp. 57–60.
- [108] Huachuan Qiu et al. “Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support”. In: *ArXiv preprint abs/2305.00450* (2023). URL: <https://arxiv.org/abs/2305.00450>.
- [109] Hannah Rashkin et al. “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 5370–5381. DOI: [10.18653/v1/P19-1534](https://doi.org/10.18653/v1/P19-1534). URL: <https://aclanthology.org/P19-1534>.

- [110] Sara Reardon. “AI Chatbots Could Help Provide Therapy, but Caution Is Needed”. In: *Scientific American* (2023). Accessed: 2024-05-13. URL: <https://www.scientificamerican.com/article/ai-chatbots-could-help-provide-therapy-but-caution-is-needed/>.
- [111] Esther Rolf et al. “Resolving label uncertainty with implicit posterior models”. In: *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*. Ed. by James Cussens and Kun Zhang. Vol. 180. Proceedings of Machine Learning Research. PMLR, 2022, pp. 1707–1717. URL: <https://proceedings.mlr.press/v180/rolf22a.html>.
- [112] Stephen Roller et al. “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, 2021, pp. 300–325. DOI: [10.18653/v1/2021.eacl-main.24](https://doi.org/10.18653/v1/2021.eacl-main.24). URL: <https://aclanthology.org/2021.eacl-main.24>.
- [113] Sahand Sabour, Chujie Zheng, and Minlie Huang. “CEM: Commonsense-Aware Empathetic Response Generation”. In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 11229–11237. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21373>.
- [114] Peter Salovey and John D Mayer. “Emotional intelligence”. In: *Imagination, cognition and personality* 9.3 (1990), pp. 185–211.

- [115] Eric F Schneider, Stephanie Gardner, and Jill T Johnson. "Development of a practical examination utilizing standardized participants for disease state management credentialing". In: *American Journal of Pharmaceutical Education* 64.2 (2000), pp. 173–176.
- [116] Iris K Schneider et al. "Let's not be indifferent about neutrality: Neutral ratings in the International Affective Picture System (IAPS) mask mixed affective responses." In: *Emotion* 16.4 (2016), p. 426.
- [117] Lifeng Shang, Zhengdong Lu, and Hang Li. "Neural Responding Machine for Short-Text Conversation". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015, pp. 1577–1586. DOI: [10.3115/v1/P15-1152](https://doi.org/10.3115/v1/P15-1152). URL: <https://aclanthology.org/P15-1152>.
- [118] Ashish Sharma et al. "A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 5263–5276. DOI: [10.18653/v1/2020.emnlp-main.425](https://doi.org/10.18653/v1/2020.emnlp-main.425). URL: <https://aclanthology.org/2020.emnlp-main.425>.
- [119] Ashish Sharma et al. "Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support". In: *Nature Machine Intelligence* 5.1 (2023), pp. 46–57.
- [120] Adam Smith. "Cognitive empathy and emotional empathy in human behavior and evolution". In: *The Psychological Record* 56.1 (2006), pp. 3–21.

- [121] Zhenqiao Song et al. “Generating Responses with a Specific Emotion in Dialog”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3685–3695. DOI: [10 . 18653 / v1 / P19 - 1359](https://doi.org/10.18653/v1/P19-1359). URL: <https://aclanthology.org/P19-1359>.
- [122] Robyn Speer, Joshua Chin, and Catherine Havasi. “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. Ed. by Satinder P. Singh and Shaul Markovitch. AAAI Press, 2017, pp. 4444–4451. URL: <http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972>.
- [123] Hao Sun et al. “PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 1489–1503. DOI: [10 . 18653 / v1 / 2021 . findings - acl . 130](https://doi.org/10.18653/v1/2021.findings-acl.130). URL: <https://aclanthology.org/2021.findings-acl.130>.
- [124] Shichao Sun and Wenjie Li. “Alleviating exposure bias via contrastive learning for abstractive text summarization”. In: *ArXiv preprint abs/2108.11846* (2021). URL: <https://arxiv.org/abs/2108.11846>.
- [125] Joshua K Swift and Jennifer L Callahan. “Decreasing treatment dropout by addressing expectations for treatment length”. In: *Psychotherapy Research* 21.2 (2011), pp. 193–200.
- [126] M Inés Torres et al. “A spoken dialogue system for the empathic virtual coach”. In: *9th International Workshop on Spoken Dialogue System Technology*. 2019, pp. 259–265.

- [127] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *ArXiv preprint abs/2302.13971* (2023). URL: <https://arxiv.org/abs/2302.13971>.
- [128] Quan Tu et al. “MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 308–319. DOI: [10.18653/v1/2022.acl-long.25](https://doi.org/10.18653/v1/2022.acl-long.25). URL: <https://aclanthology.org/2022.acl-long.25>.
- [129] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Ed. by Isabelle Guyon et al. 2017, pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [130] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. “CIDEr: Consensus-based image description evaluation”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 4566–4575. DOI: [10.1109/CVPR.2015.7299087](https://doi.org/10.1109/CVPR.2015.7299087). URL: <https://doi.org/10.1109/CVPR.2015.7299087>.
- [131] Ashwin K Vijayakumar et al. “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models”. In: (2016).
- [132] Richard S Wallace. “The anatomy of ALICE”. In: *Parsing the turing test*. Springer, 2009, pp. 181–210.

- [133] Hongru Wang et al. “Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 12047–12064.
- [134] Jiashuo Wang, Yi Cheng, and Wenjie Li. “CARE: Causality Reasoning for Empathetic Responses by Conditional Graph Generation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 729–741. URL: <https://aclanthology.org/2022.findings-emnlp.51>.
- [135] Jiashuo Wang et al. “Aligning language models with human preferences via a bayesian approach”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [136] Jiashuo Wang et al. “Empathetic response generation through graph-based multi-hop reasoning on emotional causality”. In: *Knowledge-Based Systems* 233 (2021), p. 107547.
- [137] Jiashuo Wang et al. “Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback”. In: *Findings of the Association for Computational Linguistics ACL 2024*. 2024, pp. 567–585.
- [138] Jiashuo Wang et al. “Towards a Client-Centered Assessment of LLM Therapists by Client Simulation”. In: *ArXiv preprint abs/2406.12266* (2024). URL: <https://arxiv.org/abs/2406.12266>.
- [139] Ruiyi Wang et al. “PATIENT: Using Large Language Models to Simulate Patients for Training Mental Health Professionals”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 12772–12797.

- [140] Tianming Wang and Xiaojun Wan. “T-CVAE: Transformer-Based Conditioned Variational Autoencoder for Story Completion”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, 2019, pp. 5233–5239. DOI: [10.24963/ijcai.2019/727](https://doi.org/10.24963/ijcai.2019/727). URL: <https://doi.org/10.24963/ijcai.2019/727>.
- [141] Jason Wei et al. “Finetuned Language Models are Zero-Shot Learners”. In: *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL: <https://openreview.net/forum?id=gEZrGCozdqR>.
- [142] Joseph Weizenbaum. “ELIZA—a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1 (1966), pp. 36–45.
- [143] Anuradha Welivita and Pearl Pu. “Curating a Large-Scale Motivational Interviewing Dataset Using Peer Support Forums”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 3315–3330. URL: <https://aclanthology.org/2022.coling-1.293>.
- [144] Anuradha Welivita and Pearl Pu. “Curating a Large-Scale Motivational Interviewing Dataset Using Peer Support Forums”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 3315–3330. URL: <https://aclanthology.org/2022.coling-1.293>.

- [145] Anuradha Welivita and Pearl Pu. "HEAL: A Knowledge Graph for Distress Management Conversations". In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 11459–11467. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21398>.
- [146] Anuradha Welivita and Pearl Pu. "Is chatgpt more empathetic than humans?" In: *ArXiv preprint abs/2403.05572* (2024). URL: <https://arxiv.org/abs/2403.05572>.
- [147] Anuradha Welivita, Yubo Xie, and Pearl Pu. "A Large-Scale Dataset for Empathetic Response Generation". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 1251–1264. DOI: [10.18653/v1/2021.emnlp-main.96](https://doi.org/10.18653/v1/2021.emnlp-main.96). URL: <https://aclanthology.org/2021.emnlp-main.96>.
- [148] Anna Wierzbicka. *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press, 1999.
- [149] Ka Wong, Praveen Paritosh, and Lora Aroyo. "Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability". 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)*. Online: Association for Computational Linguistics, 2021, pp. 7053–7065. DOI: [10.18653/v1/2021.acl-long.548](https://doi.org/10.18653/v1/2021.acl-long.548). URL: <https://aclanthology.org/2021.acl-long.548>.

- [150] World Health Organization. *Depressive disorder (depression)*. Accessed: 27 March 2025. Mar. 2023. URL: <https://www.who.int/news-room/fact-sheets/detail/depression>.
- [151] Zixiu Wu et al. “Anno-mi: A dataset of expert-annotated counselling dialogues”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6177–6181.
- [152] Rui Xia, Mengran Zhang, and Zixiang Ding. “RTHN: A RNN-Transformer Hierarchical Network for Emotion Cause Extraction”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, 2019, pp. 5285–5291. DOI: [10.24963/ijcai.2019/734](https://doi.org/10.24963/ijcai.2019/734). URL: <https://doi.org/10.24963/ijcai.2019/734>.
- [153] Chloe Xiang. “‘He Would Still Be Here’: Man Dies by Suicide After Talking with AI Chatbot, Widow Says”. In: *Vice* (2023). Accessed: 2024-05-14. URL: <https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says>.
- [154] Haojie Xie et al. *PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling*. 2024. URL: <https://arxiv.org/abs/2412.13660>.
- [155] Tianbao Xie et al. “UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 602–631. URL: <https://aclanthology.org/2022.emnlp-main.39>.

- [156] Ancheng Xu et al. "AutoCBT: An Autonomous Multi-agent Framework for Cognitive Behavioral Therapy in Psychological Counseling". In: *ArXiv preprint abs/2501.09426* (2025). URL: <https://arxiv.org/abs/2501.09426>.
- [157] Diyi Yang et al. "Social Skill Training with Large Language Models". In: *ArXiv preprint abs/2404.04204* (2024). URL: <https://arxiv.org/abs/2404.04204>.
- [158] Chenhao Zhang et al. "Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling". In: *ArXiv preprint abs/2405.16433* (2024). URL: <https://arxiv.org/abs/2405.16433>.
- [159] Saizheng Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2204–2213. DOI: [10.18653/v1/P18-1205](https://doi.org/10.18653/v1/P18-1205). URL: <https://aclanthology.org/P18-1205>.
- [160] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- [161] Xingxing Zhang et al. "Momentum Calibration for Text Generation". In: *ArXiv preprint abs/2212.04257* (2022). URL: <https://arxiv.org/abs/2212.04257>.

- [162] Yiqun Zhang et al. “Stickerconv: generating multimodal empathetic responses from scratch”. In: *ArXiv preprint abs/2402.01679* (2024). URL: <https://arxiv.org/abs/2402.01679>.
- [163] Zhijie Zhang, Jianmin Zheng, and Nadia Magnenat Thalmann. “Real and apparent personality prediction in human-human interaction”. In: *2022 International Conference on Cyberworlds (CW)*. IEEE, 2022, pp. 187–194.
- [164] Weixiang Zhao et al. “TransESC: Smoothing Emotional Support Conversation via Turn-Level State Transition”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 6725–6739. URL: <https://aclanthology.org/2023.findings-acl.420>.
- [165] Yao Zhao et al. “Calibrating sequence likelihood improves conditional language generation”. In: *ArXiv preprint abs/2210.00045* (2022). URL: <https://arxiv.org/abs/2210.00045>.
- [166] Peixiang Zhong, Di Wang, and Chunyan Miao. “An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss”. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 7492–7500. DOI: [10.1609/aaai.v33i01.33017492](https://doi.org/10.1609/aaai.v33i01.33017492). URL: <https://doi.org/10.1609/aaai.v33i01.33017492>.
- [167] Peixiang Zhong et al. “Towards Persona-Based Empathetic Conversational Models”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Com-

- putational Linguistics, 2020, pp. 6556–6566. DOI: [10.18653/v1/2020.emnlp-main.531](https://doi.org/10.18653/v1/2020.emnlp-main.531). URL: <https://aclanthology.org/2020.emnlp-main.531>.
- [168] Hao Zhou et al. “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 730–739. URL: <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16455>.
- [169] Jie Zhou et al. “Graph neural networks: A review of methods and applications”. In: *AI Open* 1 (2020), pp. 57–81.
- [170] Li Zhou et al. “The Design and Implementation of XiaoIce, an Empathetic Social Chatbot”. In: *Computational Linguistics* 46.1 (2020), pp. 53–93. DOI: [10.1162/coli_a_00368](https://doi.org/10.1162/coli_a_00368). URL: <https://aclanthology.org/2020.c1-1.2>.
- [171] Xianda Zhou and William Yang Wang. “MojiTalk: Generating Emotional Responses at Scale”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1128–1137. DOI: [10.18653/v1/P18-1104](https://doi.org/10.18653/v1/P18-1104). URL: <https://aclanthology.org/P18-1104>.
- [172] Caleb Ziems et al. “The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin,

Ireland: Association for Computational Linguistics, 2022, pp. 3755–3773.
DOI: [10.18653/v1/2022.acl-long.261](https://doi.org/10.18653/v1/2022.acl-long.261). URL: <https://aclanthology.org/2022.acl-long.261>.