Dialogue Comprehension and Personalization for Empathetic Response Generation

Yahui Fu

Graduate School of Informatics
Kyoto University

Abstract

Incorporating empathy into spoken dialogue systems is crucial for improving interaction with conversational robots and virtual agents, as empathy is the emotional bonding among humans; Conversational robots and virtual agents expressing empathy would give humans a feeling of being understood and satisfied with the conversation. This thesis addresses empathetic response generation for text-based dialogue systems from the perspective of appropriate dialogue comprehension and personalization.

Generally, empathy is embodied in the aspects of both contextual understanding and affective expression. However, previous studies often focus on either aspect. We first address this problem by generating appropriate empathetic responses with both aspects via modeling emotion and content consistency between the user's input and empathetic response. Moreover, it is necessary to comprehend the cause-and-effect relationships in response generation. The end-to-end generation model operates as a black box, making it unclear what factors lead to a particular response in a given context. To address this issue, we further explore causal reasoning to make the generated empathetic responses explainable. An appropriate empathetic response also depends on personality traits. Recognition of the user's personality and the development of systems that accordingly express a consistent personality are important for enhancing rapport and engagement in the interactions. To achieve this, we enhance the personality recognition in dialogue and then stylize the system to generate responses that are both empathetic and reflective of a distinct personality.

Chapter 2 provides an overview of dialogue systems, specifically emphasizing techniques for empathetic response generation (ERG).

In Chapter 3, a dual variational generative model (DVG) is proposed for empathetic response generation based on both contextual understanding and affective expression. Specifically, an emotion classifier and a variational model are incorporated into a dual response and context generative model to learn the emotion and content consistencies efficiently. DVG also uses reconstruction loss used in variational autoencoder for both contexts and responses. Evaluations on both Japanese and English EMPATHETICDIALOGUES datasets demonstrate DVG's superiority in generating empathetic responses with contextual and emotional appropriateness. In addition to the DVG model, we propose an auxiliary retrieval

system to improve empathetic response generation. Furthermore, the proposed model's ability is extended to general response generation, which is not specific to empathetic but also chitchatting dialogue systems. We evaluated our system's effectiveness in enhancing dialogue by a virtual agent. Subsequently, we integrated the system into a humanoid robot for practical application.

Chapter 4 describes the empathetic response generation based on causal reasoning. Recent approaches mainly focus on understanding the causalities of context from the user's perspective, ignoring the system's perspective. We propose a commonsense-based causality reasoning for diverse empathetic response generation that considers both the user's perspective (user's desires and reactions) and the system's perspective (system's intentions and reactions). Enhances ChatGPT's ability to reason for the system's perspective by integrating in-context learning with common-sense knowledge. Then, the commonsense-based causality explanation is integrated into both ChatGPT and a T5-based model. The integration of T5 with ChatGPT's reasoning capability realizes more empathetic responses that result in better performances. ChatGPT with the causality explanation can generate more empathetic and accurate responses.

Chapter 5 addresses the personality recognition of the user in dialog, which is useful for enhancing the ability of conversational robots and virtual agents to tailor user-adaptive responses. To address the challenge of the limited number of speakers in existing dialogue corpora, we introduce personality trait interpolation for speaker data augmentation. Moreover, a heterogeneous conversational graph neural network (HC-GNN) is incorporated to independently capture the interdependencies among interlocutors and the intradependencies within the speaker. Experimental results on the RealPersonaChat corpus demonstrate that increasing speaker diversity by data augmentation significantly improves personality recognition in both monologue and dialogue settings. The proposed HC-GNN outperforms baseline models, showcasing its effectiveness in dialogue setting.

Chapter 6 focuses on stylizing the empathetic response generation considering the system's personality. Specifically, a multi-grained prefix mechanism is designed to capture the intricate relationship between a system's personality and its empathetic expressions. Furthermore, a personality reinforcement module is designed to leverage contrastive learning to calibrate the generation model, ensuring that responses are both empathetic and reflective of a distinct personality. Automatic and human evaluations show the effectiveness of the proposed method in generating responses with enhanced empathy and personality expression.

Chapter 7 summarizes the findings of this thesis and discusses future work on adapting the system's empathetic style and personality to the user's personality in dialogue.

Acknowledgements

This dissertation was accomplished at the Speech and Audio Processing Laboratory, Graduate School of Informatics, Kyoto University. I would like to express my sincere appreciation to everyone who supported me throughout this journey.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Tatsuya Kawahara. His insightful comments were crucial for advancing this work, and this dissertation would not have been possible without his continuous support. I am deeply grateful for his dedication in assisting me with experiment analysis and revising my papers for submission. In addition, his guidance and encouragement, especially during moments of frustration due to the lack of progress, were instrumental in helping me regain confidence and persist in my research.

I also express my special thanks and appreciation to my mentor, Associate Professor Chenhui Chu, for holding seminars on speech translation and welcoming me into his laboratory meetings. The discussions were very fruitful, and his valuable advice significantly advanced my research.

I am profoundly grateful to Dr. Koji Inoue, Dr. Keiko Ochi, Dr. Divesh Lala, and Dr. Kenta Yamamoto from the spoken dialogue system group at Kawahara Lab. Their valuable advice and assistance have greatly benefited my research. Furthermore, they eased my transition to life in Kyoto by recommending and taking me to excellent restaurants and suggesting fascinating places to visit. Their support enriched my experience in Kyoto, making it more enjoyable, for which I am truly thankful. I am also deeply appreciative of Dr. Mikey Elmers from the same group, whose numerous insightful suggestions have significantly enhanced my research.

I am also grateful to Professor Shinsuke Mori and Associate Professor Yugo Murawaki for serving as members of my doctoral committee, and thanks for the advice and comments to improve the quality of this dissertation.

I wish to express my appreciation to members of the Speech and Audio Processing Lab. I am particularly grateful for the insightful comments from Professor Kazuyoshi Yoshii during our laboratory meetings and the diligent support from Ms. Mayumi Abe in handling numerous administrative matters. Many thanks go to Zi Haur Pang, Hao Shi and Yuan Gao

for their camaraderie in both academic and personal spheres. I also appreciate the support from all other members of the laboratory.

I extend my heartfelt appreciation to my internship mentor, Tianyu Zhao, and all other members for their tremendous support during my internship at rinna.

I would also like to thank Haiyue Song for being by my side during my PhD journey.

Lastly, I want to express my deepest gratitude to my family for their unwavering support, which has been my greatest source of strength and encouragement throughout my educational journey.

Table of contents

List of figures

[1] We utilize nine empathetic intents from Welivita and Pu [1], which do not strictly adhere to the definition of empathetic, including sympathizing and agreeing.

List of tables

Chapter 1

Introduction

1.1 Background

Spoken dialogue systems (SDS) have been developed for several decades and are now widely used in devices like smartphones, car navigation systems, and smart speakers. These systems offer various information services, such as making hotel or restaurant reservations, providing weather updates, and offering information about public transportation. SDS can also be integrated with virtual agents and conversational robots for services such as museum guidance [4], job interviewer [5], attentive listening [6], and chit-chatting[1].

To make human-robot interaction experiences more entertaining and natural, incorporating empathy into SDS is essential, as empathy is the emotional bond among humans; robots expressing empathy would give humans a feeling of being understood and satisfied with the conversation. Empathetic conversational robots and chatbots have drawn attention in the research community, for example, the human-like android ERICA [7], which has the ability to serve attentive listening for olderly people [6]; Nora, an empathetic dialogue system with a web-based virtual agent in the role of a psychologist [8, 9]; Chatbot CAiRE, which can detect user emotion by textual analysis and respond in an empathetic manner [10]; and a Korean multimodal empathetic dialogue system, which is able to show seven different cat face-based motions [11]. Empathetic response generation is essential for such systems. However, previous systems use template-based or simple language models for empathetic response generation, which cannot effectively express empathy.

This thesis addresses two key challenges in the text-based empathetic responses generation: dialogue comprehension and dialogue personalization. The focus is on text modality, as

[1] https://japantoday.com/category/features/new-products/palm-sized-ai-chat-bot-provides-autonomous-conversation

it is the primary medium of interaction in applications such as chatbots, virtual assistants, and online counseling, where text-based empathetic response generation is directly applicable.

1.2 Task Formulation

Empathetic response generation (ERG) is an essential task within open-domain dialogue systems, with various real-world applications, such as mental health support, customer service, and social companion systems. Unlike general response generation tasks that focus solely on coherence and relevance, ERG aims to produce responses that make sense in context and resonate emotionally with the user. The objective is to create interactions where the system understands and responds appropriately to the user's emotional state and experiences [12], fostering a sense of connection and understanding.

In the task of empathetic response generation, the input X is a t-turn dialogue received as the *context*. The output y is the upcoming $(t+1)$-th utterance, referred to as the *response* to its preceding *context*. The task is usually formulated as a sequence generation and emotion recognition problem, where the output is a variable-length sequence of tokens y and the user's emotional state e. The architecture for ERG typically comprises three key components as shown in Fig 1.1: an encoder, an emotion recognizer, and a decoder. Both the emotion recognizer and the response generator are connected to the encoder in a multitask learning manner. The encoder processes the input sequence X and converts it into a fixed-size context vector, capturing the useful features and dependencies within the input *context*. The final state of the encoder is used to predict the user's emotional state by the emotion recognizer and as the initial input into the decoder for response generation. Furthermore, with the development of large language models that leverage extensive pretraining on diverse datasets, designing prompts with zero-shot or few-shot learning using in-context learning has shown effective for ERG.

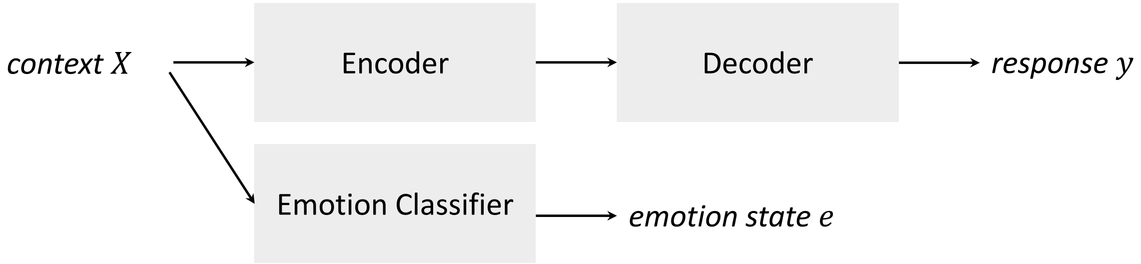


Fig. 1.1 Typical architecture of empathetic response generation.

1.3 Problems of Interest

1.3.1 Dialogue Comprehension for Empathetic Response Generation (ERG)

Empathy is the ability of humans to put themselves in another's position, encompassing an understanding of the other's experiences and feelings to respond appropriately. As shown in Fig 1.2, an empathetic response resonates with the emotion and experience of the speaker, fostering a sense of understanding and support, while non-empathetic responses lack this emotional engagement, potentially leading to feelings of detachment or indifference.

In general, empathy includes affection and cognition [12]. An incorrect emotional expression or misunderstanding the user's situation can result in a generated response that appears strange and disconnects from the user. For example, for the user's input, *"I studied so hard for 3 months straight for my bar exam,"* a response with an incorrect emotional expression may be *"I am sorry to hear that. Did you pass your exam?"* and a response with context misunderstanding may be like *"It is good to hear that you become a lawyer."* Therefore, both correct context comprehension and appropriate emotional expression are essential for generating empathetic responses that engage the interlocutor.

Recent advancements in generative models have significantly enhanced the accuracy and usability of empathetic response generation. However, generative models often function as black boxes, which results in a lack of controllability. This uncontrollability can lead to the generation of unnatural responses with grammatical or logical errors (e.g., "That is so sweet. I am sorry to hear that.") or safe but generic and meaningless responses (e.g., "I see"). Consequently, it is critically important to improve the generation of empathetic responses with explainability and controllability.

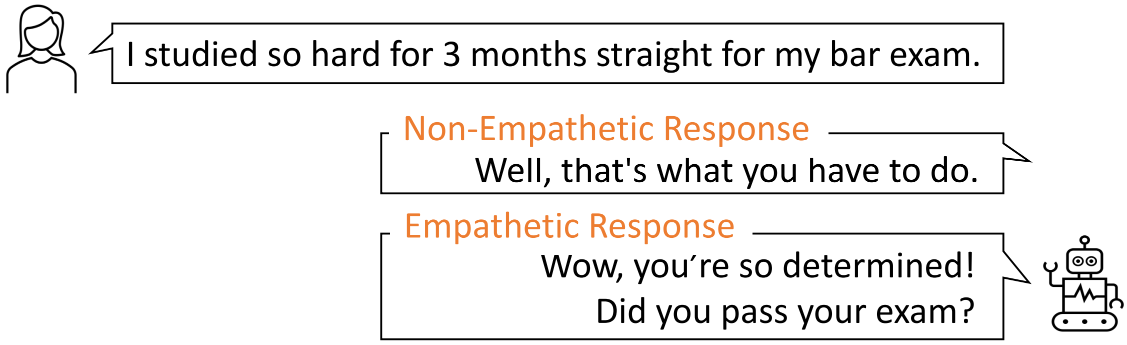

I studied so hard for 3 months straight for my bar exam.
Non-Empathetic Response
Well, that's what you have to do.
Empathetic Response
Wow, you're so determined!
Did you pass your exam?

Fig. 1.2 Example of empathetic and non-empathetic responding.

1.3.2 Dialogue Personalization for ERG

Personality recognition of users and the development of systems that express a consistent personality are crucial for enhancing the rapport of human-robot interactions. Moreover, an appropriate empathetic response also depends on the personality traits [2].

One widely-used method for achieving personality expression in systems is through the use of personas [13, 14]. This approach involves specifying detailed personal information for the system, which helps in expressing individuality and creating a consistent character that users can relate to. Another popular method is to embody the system with traits from the Big Five personality model [15] or the Myers-Briggs Type Indicator (MBTI) [16]. The Big Five model categorizes human personality into five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The MBTI framework, on the other hand, divides personality into sixteen types based on four dichotomies: extraversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving. An appropriate empathetic response may depend on the personality traits. Richendoller and Weaver III [2] indicate that individuals with different personalities exhibit distinct preferences for empathetic expression. Therefore, recognizing the user's personality and corresponding responses based on a consistent and preferred personality allows for more engaging, and empathetic interactions, which can significantly improve user satisfaction and rapport in SDS.

To this end, there are two main challenges: accurately detecting the user's personality and stylizing empathetic response generation with a consistent personality. However, the accuracy of Big Five/MBTI personality recognition is not yet good even with the assistance of large language models (LLMs) [17]. Additionally, generating empathetic responses that maintain a consistent personality lacks exploration in the research community.

1.4 Approaches

This thesis addresses methods to enhance dialogue comprehension and personalization for empathetic response generation (ERG).

1.4.1 Modeling Emotion and Content Consistency for ERG

Empathy in conversation manifests through consistent content and emotional alignment between context and response. We introduce a dual variational generative model that efficiently captures the bidirectional relationship between context and response, enhancing both contextual understanding and affective expression. To capture such a bidirectional relationship, we utilize the mutual information from the duality of response generation and context generation. Specifically, we introduce a variational model into the dual generative model to mimic the process of context/response understanding by reconstruction and utilize an emotion classifier to capture the emotion state during the conversation for affective expression. These will enhance the shared variational variables of the dual generative model with content and emotion consistencies.

To address the problems of generative models producing dull or unnatural responses, we incorporate a response retrieval module as a fallback. This module guarantees natural and empathetic responses by retrieving them from external documents. It is activated based on emotion recognition, leveraging the merits of both generative and retrieval models to improve overall response quality. It is not easy to detect emotion accurately, and false emotion detection may mislead the retrieval process. Therefore, we quantify the uncertainty of the emotion predictions as a discriminator to control the response retrieval, which means we only switch to the retrieval when the model is confident about the emotion predicted from the context.

1.4.2 Integrating Causality Reasoning for ERG

To enhance the empathetic response generation with explainability and controllability, it is necessary to incorporate both the user's perspective (exploring his/her desire and reaction) and the system's perspective (reasoning its intention and reaction to mimic humans) for empathetic response generation. To this end, we propose a commonsense-based causality explanation approach for empathetic response generation that reasons from the user's perspective to the system's perspective.

Specifically, we utilize COMET [18], which is a pre-trained GPT-2 model [19] fine-tuned on the if-then reasoning graph from ATOMIC [20], to predict user's desire and reaction. Furthermore, ChatGPT[2] has shown its efficacy in several tasks [21]. Bang et al. [22] introduced ChatGPT's potential in causal reasoning on the human-annotated explainable CAusal REasoning dataset (E-CARE) [23]. However, it is based on whether the model can make a judgment on the correct causes or effects rather than generating causality explanations. We propose to enhance ChatGPT's ability to reason the system's perspective by integrating in-context learning with commonsense knowledge. Then, we integrate the commonsense-based causality explanation with both ChatGPT and a trainable T5-based model to examine its effectiveness.

[2] https://chatgpt.com/

1.4.3 Improving User Personality Recognition in Dialogue

The lack of data is a major obstacle to personality recognition in dialogue because annotating dialogue-level data with personality information is expensive and time-consuming. Each dialogue involves two participants, and personality traits are obtained through psychology questionnaires. Thus, we investigate a data-augmentation approach. Although previous data augmentation studies focus on generating sentence-level data *invariants* [24–27] without corresponding labels, in this study, we generate both the synthetic dialogue data and corresponding synthetic personality traits through the proposed data interpolation method, which fuses two existing data points controlled by a variable *continuous ratio*.

Additionally, accurately modeling both the interdependencies between context and interlocutors, as well as the intradependencies within speakers in dialogues, remains a significant challenge. Previous homogeneous models, such as the graph attention network [28, 29], did not consider the variations in link types. Heterogeneous models like relational graph convolution networks (RGCNs) employ different relation types to model various dependencies. Moreover, they utilize shared coefficients across all relation types, which may fail to capture the unique attributes of each relation type. To address this issue, we propose a method to independently model heterogeneous conversational interactions, capturing both contextual influences and inherent personality traits.

1.4.4 Endowing System with Consistent Personality for ERG

Expressing a consistent personality is important for enhancing rapport [30]. When the system changes its personality in a single conversation, it would make the interaction feel less human-like. In addition, an appropriate empathetic response may depend on personality traits. Richendoller and Weaver III [2] examined the relationships between psychoticism, extraversion, and neuroticism and three styles of empathic intents: empathetic, perspective-taking, and sympathetic. Their findings indicate that individuals with different personalities exhibit distinct preferences for empathetic intents, inspiring our motivation to consider the system's personality traits into empathetic response generation. However, the relationship between commonly used Big Five [15] / MBTI [16] personality traits and empathy has not been fully explored.

To address this problem, we implicitly learn these connections through the prediction of both personality traits and empathetic signals in responses. Empathetic signals include empathetic intentions and empathetic communication mechanisms (ECM) [31]. Specifically, ECM includes interpretations (IP), explorations (EX), and emotional reactions (ER). Further inspired by the prefix tuning method employed by Li and Liang [32] and Liu et al. [33], we propose a multi-grained prefix encoder aimed at discerning personality traits alongside empathetic signals. Then, we propose a personality enhancement (PE) module to calibrate the generation of empathetic responses by integrating explicitly personality traits, thereby improving the empathetic and personalized expression in the generated responses.

1.5 Organization of the Thesis

The rest of this thesis is structured as below. Fig 1.3 summarizes the organization of this thesis. Fig 1.4 explains connections and distinctions among chapters. Chapter 2 provides an overview of dialogue systems, specifically emphasizing techniques for empathetic response generation (ERG). Chapter 3 presents a method that combines the strengths of both generative and retrieval systems for empathetic response generation, considering both contextual understanding and affective expression by a dual variational generative model. Chapter 4 on the comprehension of the cause-and-effect relationships in empathetic response generation. A commonsense-based causality explanation approach is proposed for empathetic response generation that reasons from the user's perspective to the system's perspective. Chapter 5 focuses on improving the personality recognition of users in dialogue. A personality trait interpolation is proposed for speaker data augmentation. Additionally, a heterogeneous conversational graph network is proposed to independently capture both contextual influences and inherent personality traits. Chapter 6 is based on the assumption that appropriate empathetic responses also depend on personality traits. We propose a multi-grained prefix encoder aimed at discerning personality traits alongside empathetic signals, and a personality enhancement (PE) module to calibrate the generation of empathetic responses by integrating explicitly personality traits.

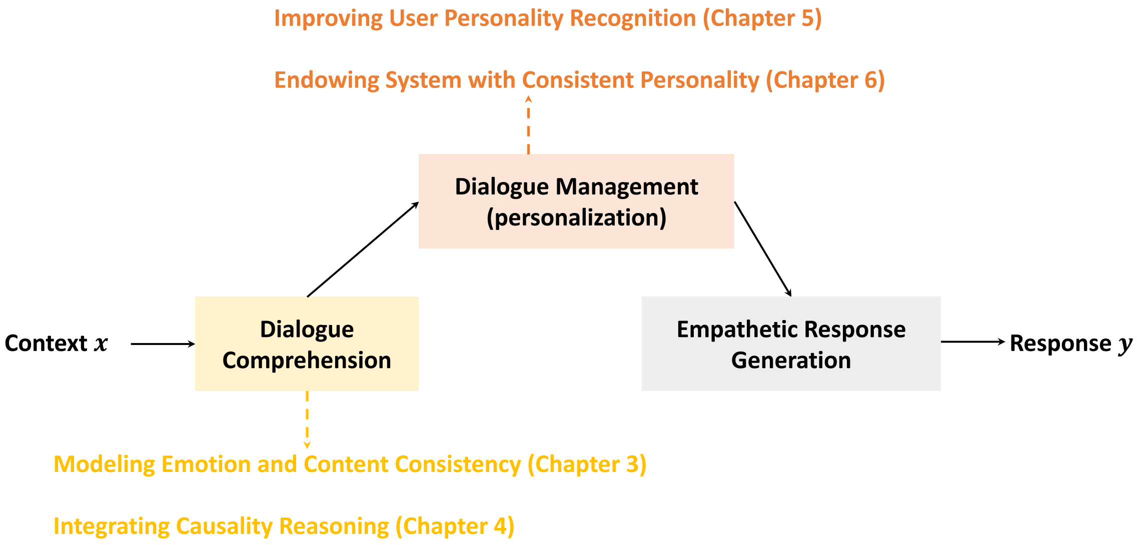

Improving User Personality Recognition (Chapter 5)
Endowing System with Consistent Personality (Chapter 6)
Dialogue Management (personalization)
Context x
Dialogue Comprehension
Empathetic Response Generation
Response y
Modeling Emotion and Content Consistency (Chapter 3)
Integrating Causality Reasoning (Chapter 4)

Fig. 1.3 Organization of the thesis.

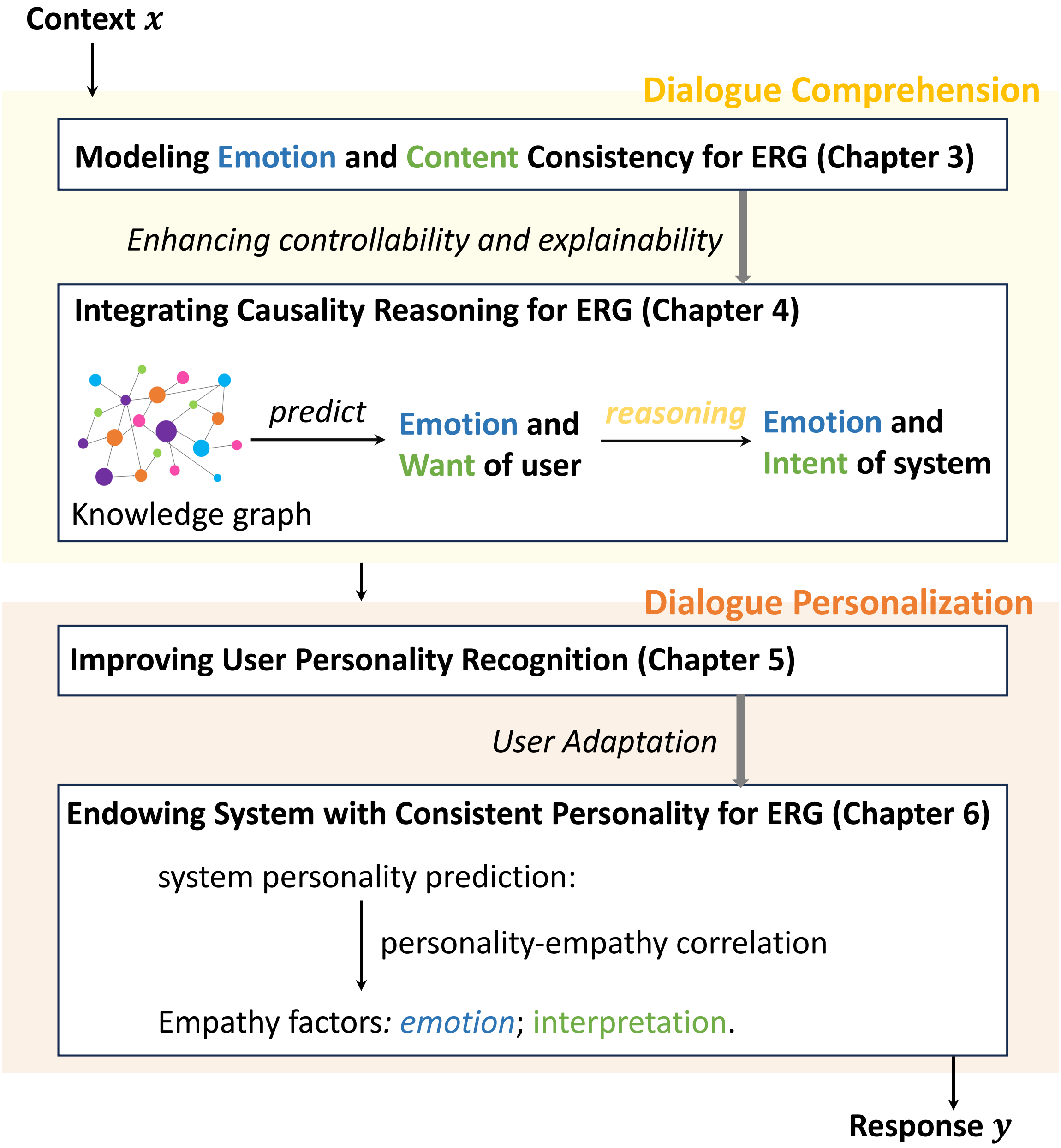

Context x
Dialogue Comprehension
Modeling Emotion and Content Consistency for ERG (Chapter 3)
Enhancing controllability and explainability
Integrating Causality Reasoning for ERG (Chapter 4)
predict
Emotion and Want of user
reasoning
Emotion and Intent of system
Knowledge graph
Dialogue Personalization
Improving User Personality Recognition (Chapter 5)
User Adaptation
Endowing System with Consistent Personality for ERG (Chapter 6)
system personality prediction:
personality-empathy correlation
Empathy factors: emotion; interpretation.
Response y

Fig. 1.4 Connections and distinctions among chapters.

Chapter 2

Literature Review

This chapter provides an overview of dialogue systems, focusing on techniques for empathetic response generation (ERG). Section 2.1 introduces a brief history of dialogue systems. Section 2.2 introduces the basic models that will be used in later chapters. Finally, Section 2.3 offers a comprehensive review of the ERG task.

2.1 A Brief History of Dialogue Systems

Early Rule-based and Statistic-based Dialogue Systems

The early-stage dialogue system can be traced back to the 1960s with the development of *ELIZA* [34]. *ELIZA* simulated a psychotherapist using simple pattern-matching and substitution methodologies. *ELIZA* generated responses by identifying key patterns in user input and applying pre-defined transformation rules. In the 1970s, systems like *PARRY* emerged, designed to simulate the thought processes of a paranoid schizophrenic, utilizing more complex rules and emotional models [35]. *SHRDLU* [36] was designed to understand and respond to natural language within a simulated "blocks world," allowing it to execute commands, answer questions, and manipulate objects such as blocks, pyramids, and cubes.

Since the 1990s, numerous **task-oriented dialogue systems** have been developed to assist users in accomplishing specific tasks such as planning routes, booking flights, or finding restaurants. *VOYAGER* [37], an urban exploration and navigation system, allowed users to interact through spoken dialogue, text, and graphics. *TRAINS* [38, 39] was designed to assist managers in solving routing problems within a transportation domain, using a map that displayed cities and rail connections. *Airline Travel Information System (ATIS)* [40] provided flight information. A large-amount data collection allowed for invention of statistical methods, which greatly improved the performance of task-oriented systems.

Levin et al. [41] introduced the use of Markov Decision Process (MDP) and reinforcement learning algorithms to optimize dialogue strategies, marking a significant step toward data-driven approaches in dialogue system design. Williams and Young [42] modeled a spoken dialogue system as a Partially Observable Markov Decision Process (POMDP), enhancing their robustness in handling uncertainties caused by speech recognition errors. Gasic et al. [43] applied Gaussian Processes to POMDP-based dialogue managers, greatly enhancing the efficiency of policy optimization. This approach was applied to the *tourist information system for Cambridge*, enabling users to inquire about restaurants, hotels, museums, and other local attractions.

An intelligent conversational agent should not be confined to a specific task; instead, it should possess the ability to engage in dialogue across various domains. Such a system is known as an **open-domain dialogue system**. Some spoken dialogue systems in smartphone applications, such as *Siri*, integrate two approaches: a task-oriented system that relies on well-defined domain knowledge, often using a relational database (RDB) or information retrieval techniques like named-entity recognition, and an open-domain system that utilizes pattern matching for generating simple chitchat responses.

Neural Dialogue Systems

The development of dialogue systems has greatly benefited from the advent of deep learning. Task-oriented dialogue systems have been significantly improved by substituting statistical models with neural networks. *MultiWOZ* [44] set new benchmarks by introducing multi-domain dialogues and more complex interactions, evolving the development of end-to-end neural architectures. Fig 2.1 shows an example of this type of system.

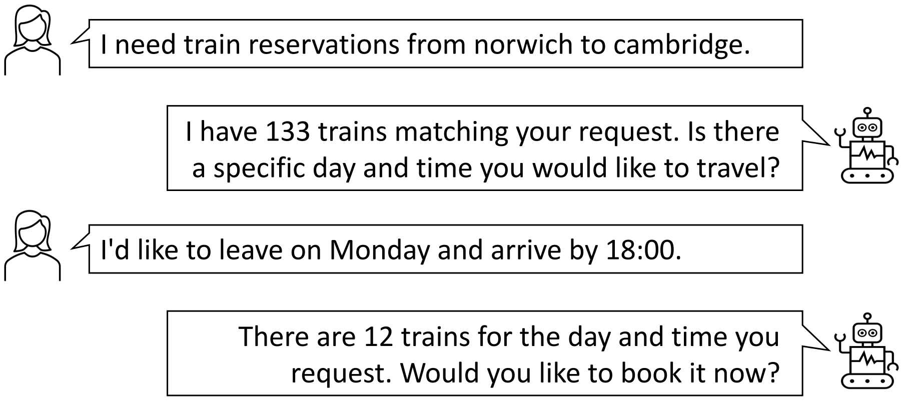

I need train reservations from norwich to cambridge.
I have 133 trains matching your request. Is there a specific day and time you would like to travel?
I'd like to leave on Monday and arrive by 18:00.
There are 12 trains for the day and time you request. Would you like to book it now?

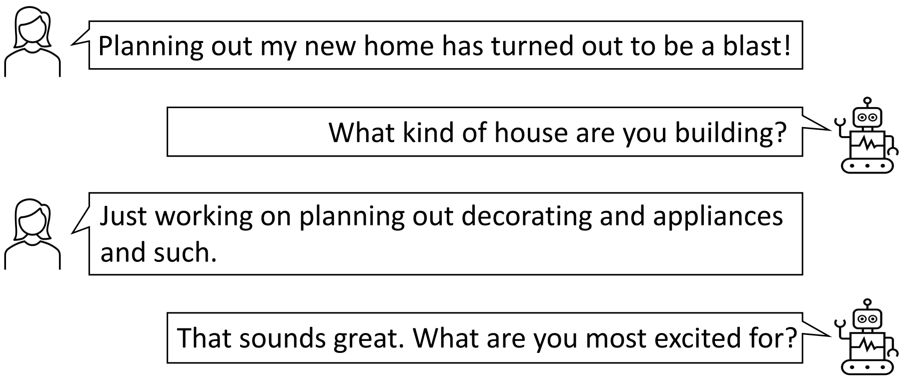
Fig. 2.1 Task-oriented dialogue-example from the MultiWOZ dataset.

Open-domain dialogue systems, on the other hand, have benefited from a sequence-to-sequence (seq2seq) learning paradigm. Sutskever et al. [45] introduced *Sequence-to-Sequence (Seq2Seq)* models using a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector, demonstrating early success in machine translation. Bahdanau et al. [46] introduced the *attention mechanism*, which allows models to focus on the relevant parts of the input sequence when generating each word in the output sequence. This significantly improved the performance of *Seq2Seq* models by enhancing their ability to handle long and complex sentences. Vaswani et al. [47] proposed the *Transformer* architecture, which based solely on attention mechanisms, dispensing with the recurrence and convolutions used in previous models. *Transformer* significantly improved training efficiency and model performance, becoming the foundation for both pretrained language models, such as Bidirectional Encoder Representations from Transformers (BERT) [48], Generative Pretrained Transformer (GPT) [49], RoBERTa [50], BART [51], DialoGPT [52], T5 [53], and BlenderBot [54], as well as other large language models (LLMs), such as GPT-3/3.5/4o, LLaMA [55], Mistral [56], and Vicuna [57]. Prompting empowers LLM with great potential besides traditional seq2seq or finetuning methods. In-context learning with a few shots [58] allows the model to quickly adapt to a new task or a new domain by showing several examples that are predetermined or retrieved from a dataset by similarity. Chain-of-thought prompting [59] enables reasoning ability, which is crucial for complex tasks or samples.

Stylistic Open-domain Dialogue Systems

Stylistic open-domain dialogue systems focus on generating responses with specific styles by using neural networks, such as incorporating persona or empathy. These systems aim to enhance the user experience by making interactions more human-like and engaging. This is achieved by tailoring the dialogue system's responses to align with a defined persona, personality, or empathetic tone. Fig 2.2 shows an example of such systems.

Some efforts in this area focused on incorporating persona profiles into responses. For instance, PERSONA-CHAT introduced the concept of conditioning responses on a predefined persona to make interactions more consistent and relevant [13]. Some conversational systems have been developed to generate empathetic responses based on the benchmark dataset EMPATHETICDIALOGUES [60]. These systems aim to produce responses that reflect understanding of the user's emotions and experiences, thereby improving the user's satisfaction and engagement. Adding personality or empathy to response generation aims to create a more natural and engaging conversational experience, fostering a sense of connection and understanding between the user and the system. The commonly used neural networks for stylistic response



Planning out my new home has turned out to be a blast!

What kind of house are you building?

Just working on planning out decorating and appliances and such.

That sounds great. What are you most excited for?

Fig. 2.2 Open-domain dialogue-example from the EMPATHETICDIALOGUES dataset.

generation can also be categorized into three approaches: (1) training a vanilla Transformer encoder-decoder from scratch using a target dataset; (2) using pretrained language models (*PLMs*) that have been trained on a large dataset of text data in a self-supervised manner before being finetuned on the target dataset; (3) utilizing large language models (*LLMs*) that leverage extensive pretraining on diverse datasets, followed by reinforcement learning with human feedback (RLHF) to align with human preferences.

2.2 Model Basics

This section formulates the commonly-used models for generating responses in open-domain dialogue systems.

2.2.1 HRED: Hierarchical Recurrent Encoder Decoder

The seq2seq model enables the generation of responses to single-utterance contexts. To utilize multi-utterance contexts, the Hierarchical Recurrent Encoder-Decoder (HRED) model [61] employs a hierarchical structure that captures dependencies at both the utterance and conversation levels. The HRED model consists of three main components: the encoder RNN, the context recurrent neural network (RNN), and the decoder. Each of these components plays a specific role in the processing of the dialogue data.

Encoder RNN: The encoder processes individual utterances within a dialogue. Given an utterance $\mathbf{u}_i = \{w_{i,1}, w_{i,2}, \dots, w_{i,T_i}\}$, where $w_{i,t}$ represents the t-th word in the i-th utterance, the encoder RNN maps each utterance to an utterance vector $\mathbf{h}_{i,t}$:

$$\mathbf{h}_{i,t} = \mathrm{RNN}_{\mathrm{enc}}(\mathbf{h}_{i,t-1}, \mathbf{w}_{i,t}). \tag{2.1}$$

The final hidden state $\mathbf{h}_{i,T_i}$ represents the encoded representation of the utterance $\mathbf{u}_i$. *Context RNN*: The context RNN models the dependencies between consecutive utterances in a dialogue. Given a context $\mathbf{U} = \{u_1, u_2, \ldots, u_i\}$, it takes the encoded representation $\mathbf{h}_{i,T_i}$ from the encoder and updates its hidden state $\mathbf{s}_i$:

$$\mathbf{s}_i = \text{RNN}_{\text{ctx}}(\mathbf{s}_{i-1}, \mathbf{h}_{i,T_i}). \tag{2.2}$$

Decoder RNN: The decoder generates the next utterance $\mathbf{u}_{i+1}$ based on the context RNN's hidden state $\mathbf{s}_i$. The decoder RNN predicts each word $w_{i+1,t}$ sequentially.

$$\mathbf{h}_{i+1,t} = \text{RNN}_{\text{dec}}(\mathbf{h}_{i+1,t-1}, \mathbf{w}_{i+1,t-1}, \mathbf{s}_i). \tag{2.3}$$

The probability of generating the next word $w_{i+1,t}$ is given by:

$$P(w_{i+1,t}|w_{i+1,<t}, \mathbf{s}_i) = \text{softmax}(\mathbf{W}_o \mathbf{h}_{i+1,t} + \mathbf{b}_o), \tag{2.4}$$

where softmax layer converts $\mathbf{h}_{i+1,t}$ into a probability distribution over the token vocabulary.

2.2.2 Decoding Algorithms

The decoding algorithm decides how to transform $P(w_{i+1,t}|w_{i+1,<t}, \mathbf{s}_i)$ to the generated token $w_{i+1,t}$. The algorithm takes the probability distribution of each token in the target vocabulary $P(w_{i+1,t}|w_{i+1,<t}, \mathbf{s}_i)$ as input. We briefly describe four primary strategies used in this thesis: greedy search, beam search, top-k sampling, and top-p sampling:

(1) The greedy search selects the token with the highest conditional probability over $\mathbf{p}_t$ at each step. Specifically, let $y_{1:t-1}$ represent the sequence of tokens generated up to step $t-1$. At step t, the next token y_t is chosen as follows:

$$y_t = \arg\max_{y_t^{'}} P(y_t^{'}|y_{1:t-1}). \tag{2.5}$$

This process repeats until the end-of-sequence token (<EOS>) is generated or a predefined maximum length is reached.

(2) Beam search considers multiple potential tokens at each step by keeping track of a fixed number ("beams"). Let B be the beam size. At each step, it expands each of the B sequences by considering all possible next tokens. For each partial sequence, it calculates the total score (the sum of probabilities) for each possible extension and keeps only the B

sequences with the highest total scores:

$$B_t = \text{Top-}B\left\{P(y_{1:t}) \mid y_{1:t} \in B_{t-1} \times V\right\}, \tag{2.6}$$

where V is the vocabulary size.

(3) Top-k Sampling introduces randomness into the decoding process by selecting the next token from the top k tokens over $\mathbf{p}_t$, rather than always choosing the highest-probability token. At each step t:

$$P'(y_t) = \frac{P(y_t \mid y_{1:t-1})}{\sum_{y \in V_k} P(y \mid y_{1:t-1})}, \tag{2.7}$$

where V_k is the set of top k tokens. Top-k Sampling helps to introduce diversity into the generated sequences but may occasionally result in less coherent outputs.

(4) Top-p Sampling, also known as Nucleus Sampling, dynamically selects tokens from the smallest possible set of top tokens whose cumulative probability exceeds a threshold p. At each step t:

$$P'(y_t) = \frac{P(y_t \mid y_{1:t-1})}{\sum_{y \in V_p} P(y \mid y_{1:t-1})}, \tag{2.8}$$

where V_p is the set of tokens forming the cumulative probability p. Top-p Sampling strikes a balance between coherence and diversity.

2.2.3 Transformer

RNNs compute along the symbol positions of input and output sequences, generating hidden states $\mathbf{h}_{i,t}$ based on the previous hidden state $\mathbf{h}_{i,t-1}$ and the input at position t. This sequential processing hinders parallelization within training examples, which is problematic for long sequences due to memory constraints on batching. While HRED uses RNNs to capture dependencies in the input and context, the Transformer [47] relies on self-attention mechanisms. This allows the Transformer to handle long-range dependencies more effectively. The transformer consists of an encoder and a decoder, both composed of multiple layers of self-attention (SAN) and feed-forward neural networks (FFN), both wrapped by residual connections [62] and layer normalization. A key component of SAN sub-layers is a multi-head attention (MHA) mechanism, which allows the model to focus on different parts of the input sequence simultaneously. In the MHA, key (K), value (V), and query (Q) metrics are split into h heads with a dimension $d_k = d_{model}/h$ after linear transformations and each head performs a scaled-dot attention mechanism as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.9}$$

The combined attention output is then processed by the FFN dub-layer. The decoder generates the target sequence based on the encoded input representations and previously generated tokens. The target sequence tokens are embedded and positional encodings are added. The decoder uses masked self-attention to prevent attending to future tokens. It also attends to the encoder's output through another multi-head attention mechanism. A FFN sub-layer then processes this combined information. Finally, the output layer generates the probability distribution over the vocabulary, producing the next word in the sequence.

2.2.4 Pretrained Language Models (PLMs)

The Transformer architecture fosters the development of pretrained language models that learn knowledge through pretraining on unlabeled data across various pretraining tasks and are subsequently finetuned using labeled data from the downstream tasks.

Bidirectional Encoder Representations from Transformers: BERT

BERT [48] is based on the Transformer encoder architecture. Unlike the traditional Transformer, which processes the input in a unidirectional manner, BERT is bidirectional, taking into account both the left and right contexts of a word simultaneously. In addition, a special classification token (`[CLS]`) is added at the beginning of every sequence. The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. Sentence pairs are packed together into a single sequence and separated by a special token (`[SEP]`). A learned embedding (segment token) is added to each token to indicate whether it belongs to sentence `A` or sentence `B`. Given a sequence of tokens $\mathbf{x} = \{[CLS], x_1, x_2, [SEP], x_3, \ldots, x_n, [SEP]\}$, the input embeddings $\mathbf{E}_i$ for each token x_i are computed as:

$$\mathbf{E}_i = \mathbf{W}_{\text{token}}(x_i) + \mathbf{W}_{\text{segment}}(s_i) + \mathbf{W}_{\text{position}}(p_i), \tag{2.10}$$

where $\mathbf{W}_{\text{token}}$, $\mathbf{W}_{\text{segment}}$, and $\mathbf{W}_{\text{position}}$ are the token, segment, and position embedding matrices respectively, and s_i and p_i represent the segment and position indices for token x_i. For the training strategies, BERT is pretrained using two unsupervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP): (1) MLM: some tokens in an input are randomly masked, and the objective is to predict these masked tokens based on their context. (2) NSP: the model is given pairs of sentences and tasked with predicting whether the second sentence follows the first in the original text. After pretraining, BERT can be finetuned on specific tasks by adding a task-specific output layer. During finetuning, all parameters from BERT, as well as the additional output layer, are updated jointly to optimize the task-specific objective.

Generative Pretrained Transformer: GPT

GPT [49] employs a multi-layer *Transformer decoder* [63] architecture. It utilizes unsupervised learning to improve language understanding, demonstrating that *generative pretraining* on a diverse corpus of unlabeled text, followed by *discriminative finetuning* on specific tasks, is effective across various language tasks. Unsupervised pretraining is the first phase of training GPT, where the model learns to predict the next word in a sequence without requiring labeled data. Given an unsupervised corpus of tokens $U = \{u_1, ..., u_n\}$, the objective during pretraining is to maximize the likelihood of each token given its preceding tokens:

$$L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta), \tag{2.11}$$

where k is the context window size. θ denotes the model parameters. This model applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens:

$$\begin{aligned} h_0 &= UW_e + W_p, \\ h_l &= \text{transfomer\_block}(h_{l-1}) \forall i \in [1, n], \\ P_u &= \text{softmax}(h_n W_e^T), \end{aligned} \tag{2.12}$$

where n is the number of layers. W_e is the token embedding matrix, and W_p is the position embedding matrix. After the unsupervised pretraining phase, GPT undergoes supervised finetuning to adjust the pretrained model parameters to optimize performance for the specific supervised task. Given a labeled sequence $C = x^1, ..., x^m$, along with a label y. The inputs are passed through the pretrained model to obtain the final transformer block's activation h_l^m, which is then fed into an added linear output layer with parameters W_y to predict y:

$$\begin{aligned} P(y|x^1, ..., x^m) &= \text{softmax}(h_l^m W_y), \\ L_2(C) &= -\sum_{(x,y)} \log P(y|x_1, \ldots, x_m). \end{aligned} \tag{2.13}$$

To summarize, the training process of GPT involves two key objectives:

$$L_3(C) = L_2(C) + \lambda * L_1(C). \tag{2.14}$$

Overall, the only extra parameters that are required during finetuning are W_y, and embeddings for delimiter tokens (initialized start and end tokens of the input (`<s>`, `<e>`).

Dialogue Generative Pretrained Transformer: DialoGPT

GPT training involves two stages: unsupervised pretraining and supervised fine-tuning on target tasks; GPT-2 [19] advances this approach by utilizing a significantly larger dataset and applying unsupervised pretraining models for supervised tasks with the speculation that a good language model could do unsupervised multitask learning while training. The objective is to model $P(\text{output}|\text{input},\text{task})$. GPT-2 shows good performance on downstream tasks in a zero-shot setting – without any parameter or architecture modification.

DialoGPT inherits from GPT-2, a 12-to-48-layer Transformer with layer normalization. Like GPT-2, DialoGPT is formulated as an autoregressive (AR) language model. Unlike GPT-2, which is trained on a dataset comprising millions of webpages known as WebText for text generation, DialoGPT extends GPT-2 to address the challenges of conversational neural response generation. It achieves this by training on large-scale dialogue pairs and sessions extracted from Reddit discussion chains, enabling DialoGPT to capture the joint distribution of $P(\text{Target},\text{Source})$ in conversational flow with finer granularity.

BlenderBot

Unlike DialoGPT, BlenderBot is trained on a diverse set of dialogue tasks simultaneously. This multi-task training approach helps the model learn to balance different conversational objectives. Specifically, BlenderBot was first pretrained on the pushshift.io Reddit [64] to learn general language representations. Then it was finetuned in the ConvAI2 dataset [13] focuses on personality and engaging the other speaker, EmpatheticDialogues [65] focuses on empathy, and Wizard of Wikipedia [66] focuses on knowledge, and Blended Skill Talk [67] provides a dataset that focuses on blending these skills.

While DialoGPT employs a multilayer Transformer decoder architecture, BlenderBot employs three types of architecture: retrieval, generative, and retrieve-and-refine models, all based on Transformers. (1) Retrieval: Given a dialogue history (context) as input, this architecture uses the poly-encoder [68] to select the next dialogue utterance by scoring a large set of candidate responses and outputting the highest-scoring one. (2) Generative: This is a standard Seq2Seq Transformer architecture that generates responses rather than retrieving them from a fixed set. (3) Retrieve and refine: This approach combines a retrieval step before generation, replacing the retrieved response with the gold response $\alpha\%$ of the time. The parameter α is a hyperparameter that can be tuned, allowing for a smooth transition between retrieval and generative systems.

Text-to-Text Transfer Transformer: T5

Transfer learning, where a model is first pretrained on a data-rich task and then finetuned on a downstream task, has become a powerful technique in natural language processing (NLP). This effectiveness has led to a variety of approaches, methodologies, and practices. T5 [53] is designed to handle a wide range of NLP tasks by converting each task into a text-to-text format. T5 aims to unify various NLP problems, such as translation, summarization, and question answering, under a single framework. This unified approach improves generalization and allows the model to leverage knowledge from different tasks to improve dialogue response quality. Meanwhile, T5 is pretrained on a massive and diverse corpus of text, enabling it to understand and generate coherent text across various domains. This extensive pretraining helps T5 generate more contextually appropriate and varied responses in dialogues.

Unlike BlenderBot, T5 uses a pretraining objective similar to masked language modeling (MLM) used in BERT and then finetuned on specific downstream tasks. The pretraining involves predicting missing tokens in a text, and the finetuning involves adjusting the model for specific tasks by providing task-specific input-output pairs.

2.3 Review on Empathetic Response Generation (ERG)

This section will introduce the literature reviews on ERG task.

Non-verbal ERG

Facial expression and head gesture mirroring are common forms of empathic conveyance, often involving head nodding, laughing, eyebrow raising, and smiling [69]. For example, Hegel et al. [70] conducted a study with an anthropomorphic robot that recognizes the user's emotional state through speech intonation and then mirrors the inferred state using a corresponding facial expression. Similarly, Riek et al. [69] utilized a robot designed as a chimpanzee head to mimic users' mouth and head movements, enabling human-robot rapport. Furthermore, Jo et al. [71] and Inoue et al. [72] highlighted the importance of incorporating appropriate laughter in conversational robots to enhance empathetic interactions.

Affection-Driven ERG

Affective empathy, also known as emotional empathy, is the ability to feel what someone else feels, often described as "your pain in my heart." Affection-driven methods emphasize the affective aspect of empathy expression, detecting and leveraging the user's emotions using various structures. For example, Lin et al. [73] softly combined the output of multiple emotion-specific decoders to improve the generation of appropriate empathetic responses. Majumder et al. [74] argued that empathetic responses often mimic the speaker's emotion, then proposed emotion grouping and emotion mimicry to generate empathetic and various responses. Fu et al. [75] proposed an emotion correlation-enhanced framework for empathetic dialogue generation, which comprehensively captures emotion interactions by employing a multi-resolution emotion graph to model context-based emotion correlations.

Cognition-Driven ERG

Cognitive empathy refers to the ability to understand another person's point of view or perspective, often described as "putting oneself in others' shoes." Cognition-driven methods aim to enhance contextual understanding through different mechanisms, including exploration of *empathetic intentions* [1], *emotion cause reasoning* [76–78], *integration of common sense knowledge* [79, 80], and *additional retrieval processes* [81, 82]. For example, Kim et al. [76] extracted emotion causes from the dialogue context by utilizing a rational speech act framework. Additionally, Wang et al. [77] employs a cause-effect graph [83] to reason about the emotion causes and effects, thereby improving context understanding. Sabour et al. [84] leveraged ATOMIC [20], which is a knowledge base of commonsense reasoning inferences about if-then events to improve contextual understanding in the dialog. Fu et al. [82] adopt the retrieval system as a fallback to the generation model based on emotion classification to alleviate the difficulty of empathetic response generation.

Affection and Cognition-Driven ERG

Affection and cognition-driven methods consider both emotional and cognitive aspects in empathy expression. Sharma et al. [85] introduced three types of communication mechanisms—emotional reactions, interpretations, and explorations, representing higher-level and abstract factors in empathy expression. Additionally, Fu et al. [81] proposed a dual variational generative model that efficiently learns the bidirectional relationship between context and response in conversations, facilitating both contextual understanding and affective expression. Yang et al. [86] proposed to dynamically learn the emotion-semantic correlations through the interaction of context and emotions for empathetic response generation. Majumder et al. [87] utilized the T5 encoder-decoder model to integrate emotional presence, interpretation, and exploration for empathetic response generation.

Large Language Models (LLMs)-based Methods

With the development of LLMs such as GPT-series (like GPT-3 [58], ChatGPT, GPT4 [88]), many studies have shown their ability on various NLP tasks with either a few-shot or zero-shot setting [21, 89, 90]. Lee et al. [90] introduced two selection methods that choose in-context examples based on emotion and situation information to generate empathetic responses by GPT-3. Zhao et al. [21] showed ChatGPT's ability on empathetic response generation, Fu et al. [91] enhanced ChatGPT's ability to reason for the system's perspective by integrating in-context learning with commonsense knowledge for ERG. Cai et al. [92] leveraged the reasoning capabilities of ChatGPT to generate commonsense causal sentences through in-context learning; additionally, they modeled the emotional flow within dialogues to predict future emotional states for ERG.

Persona-based ERG

Recent studies have increasingly emphasized the integration of persona into empathetic response generation. Persona is highly correlated with personality which in turn influences empathy [2]. This integration is crucial for generating consistent responses across multi-turn conversations, which significantly enhances user engagement and satisfaction. Recent advancements in persona-based empathetic response generation fall into two distinct categories: (1) customization using explicit system-specific profiles or descriptive persona sentences [13], and (2) generation based on explicit personality traits, such as Big Five and MBTI. Specifically, Zhong et al. [14] constructed a multi-turn Persona-based Empathetic Conversation (PEC) dataset obtained from the social media Reddit. Huang et al. [93] proposed a Transformer-based architecture that incorporates retrieval-augmented prompt learning to generate persona-aware empathetic responses based on the PEC dataset. Cai et al. [94] proposed to perceive speakers' implicit emotional information by dynamically capturing persona information from the PERSONA-CHAT dataset [13] and reasoning about future emotional reactions, and incorporating them into the process of empathetic response generation. Fu et al. [95] proposed a multi-grained prefix mechanism to learn the relationship between a system's personality and its empathetic expressions, and a personality reinforcement module to calibrate the generation model to generate responses that are both empathetic and reflective of a distinct personality.

Multi-Party ERG

Zhu et al. [96] introduced the task of multi-party empathetic dialogue generation, expanding the scope of empathetic response generation to conversations involving multiple participants.

They proposed to model multi-party dialogues by constructing a dynamic graph network with temporal information and exploring participants' dynamic emotions and static sensibilities by fusing speaker information.

Multi-Modal ERG

Tavabi et al. [97] demonstrated that, beyond text, the emotional tone in language and facial expressions are strong indicators of sentiment in conversations that necessitate an empathetic response. To this end, they developed a multi-modal deep neural network designed to identify instances where an agent should express positive or negative empathetic responses. This model leverages audio, video, and language data from human-agent interactions conducted in a wizard-of-Oz setting. Fei et al. [98] introduced an avatar-based multi-modal empathetic chatbot that integrates text, sound, and vision, leveraging advancements in large language models combined with multi-modal encoders and generators. Additionally, Shen et al. [99] introduced a new multi-modal dataset, EMPATHICSTORIES++, designed to study empathy during personal experience sharing.

2.4 Conversational Datasets

This section will introduce the datasets that will be used in later chapters.

The EMPATHETICDIALOGUES dataset [60][1] comprises 24,850 open-domain, multi-turn conversations (4-8 utterances) in English between two interlocutors. Each conversation is grounded in a scenario, where one participant (*Speaker*) describes a situation associated with a specific emotion label. The dataset features 32 emotion labels, converging a broad range of positive and negative emotions. The *Speaker* initiates the dialogue by discussing their situation, while the second participant (*Listener*) interprets the scenario through the *Speaker*'s descriptions and responds accordingly.

The Japanese EMPATHETICDIALOGUES dataset [100][2] was developed by adapting the original English version [60] into Japanese. This adaptation involved translating the 32 emotion-related English words into Japanese, which were then used by Japanese speakers to construct situational sentences and dialogues. Unlike the original English version, where dialogues were conducted between two participants, the Japanese version consists of pseudo-dialogues generated by a single crowdworker. The crowdworker refers to the translated list of emotions and creates a context sentence of 1-3 sentences based on the emotions and a text dialogue of four utterances by two persons (*Speaker* and *Listener*) who interact in the

[1] https://huggingface.co/datasets/empathetic_dialogues

[2] https://github.com/nttcslab/japanese-dialog-transformers

context. In total, the dataset comprises 20,000 dialogues and 80,000 utterance pairs, with 32 evenly distributed emotion labels.

For both English and Japanese EMPATHETICDIALOGUES datasets, we train and evaluate models in later chapters to generate *Listener* responses at each conversational turn. These models are conditioned on the evolving context, with the *Speaker*'s inputs being extended incrementally to inform the *Listener*'s responses.

The PERSONA-CHAT dataset [13][3] is designed such that each speaker is characterized by a set of five profile sentences that define their persona. Conversations are conducted based on these predefined profile sets, and the resulting dialogues are collected in a pseudo-dialogue manner. The Japanese version of this dataset [100][3] adapts the original by creating a corresponding set of Japanese profile sentences and collecting conversations from Japanese speakers. In total, the Japanese `PersonaChat` dataset includes 61,794 utterances across 5,000 dialogues.

The RealPersonaChat [101][4] dataset comprising 14,000 Japanese dialogues and a total of 421,203 utterances. In this dataset, 233 participants completed a questionnaire regarding their Big Five personality score (in a range from 1-7) and then engaged in unstructured conversation. We normalize the score to 0-1.

[3] https://parl.ai/projects/personachat/

[4] https://github.com/nu-dialogue/real-persona-chat

Chapter 3

Modeling Emotion and Content Consistency For Empathetic Response Generation

3.1 Introduction

Incorporating empathy into the dialogue system is essential for improving human-robot interaction experiences, as empathy is the emotional bonding among humans; robots expressing empathy would give humans a feeling of being understood and satisfied with the conversation. For example, McNeill and Kennington [102] proposed a multimodal emotion recognition model to predict human interpretations of social robots' emotions. Winata et al. [8, 9] introduced Nora, an empathetic dialogue system with a web-based virtual agent in the role of a psychologist; Jung et al. [11] introduced a Korean multimodal empathetic dialogue system with a virtual agent. These previous systems relied on coarse-grained emotion recognition and basic response generation. In this study, we build a fine-grained empathetic system capable of offering diverse emotions and responses. Subsequently, we evaluate this system by the virtual agent Gene [103] to show vibrant empathy. The system is further implemented in the humanoid robot CommU for practical application.

In general, empathy includes aspects of contextual understanding and affection [12], which represent perceiving the user's situation and expressing emotion, such as the "Empathetic response" shown in Fig. 3.1. However, previous studies either focused on detecting user emotion and embedding emotional traits to generate responses with affection [73, 74, 104], or focused on integrating commonsense knowledge to help contextual understanding [84].

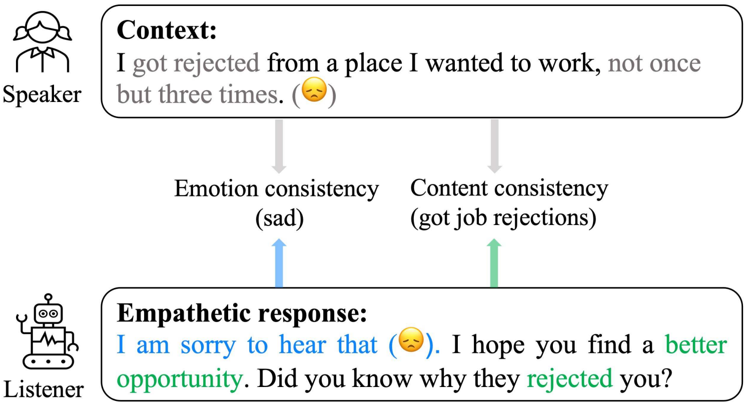

Speaker
Context:
I got rejected from a place I wanted to work, not once but three times. (😞)
Emotion consistency (sad)
Content consistency (got job rejections)
Listener
Empathetic response:
I am sorry to hear that (😞). I hope you find a better opportunity. Did you know why they rejected you?

Fig. 3.1 An example of an empathetic response from the Japanese EMPATHETICDIALOGUES dataset. Blue highlighted text denotes the affective expression, and green text implicates context understanding.

To make a further exploration on both aspects for empathetic response generation, in this paper, we propose a dual variational generative (DVG) model.

Our DVG model is based on the assumption that there exist emotion and content consistencies between context and appropriate empathetic response, as shown in Fig. 3.1. To capture such consistency, we utilize the mutual information from the duality of response generation and context generation. Specifically, (1) we introduce a variational autoencoder (VAE) into the dual generative model to mimic the process of context/response understanding by reconstruction and (2) utilize an emotion classifier to capture the emotion state during the conversation for affective expression. These will enhance the shared variational variables of the dual generative model with content and emotion consistencies.

The generative models can produce an empathetic response, but they encounter the problem of generating dull responses (generic and meaningless, such as ‘I see’) or unnatural responses (have grammatical or logical errors, such as ‘that is so sweet. I am sorry to hear that’). Instead, the retrieval-based models are guaranteed to produce natural and empathetic responses, as they are retrieved from external documents, but encounter the problem of producing responses that are not closely relevant to the dialog context. Therefore, we incorporate a response retrieval model as a fallback to the generative model based on emotion recognition to leverage the merits of both the generative and retrieval model. Specifically, we define 82 empathetic responses conditioned on 32 kinds of emotions as a controllable retrieval set. It is difficult to detect emotion accurately, and false emotion detection may mislead the retrieval process. Therefore, we quantify the uncertainty in the emotion predictions and use it as a discriminator to control the response retrieval, which means we only switch to the retrieval when the model is confident about the emotion predicted from the context.

In daily-style conversation, an empathetic response is just one kind of conversation reply, while neutral chatting also accounts for a large percentage. Therefore, we further enrich our model's ability to build a general Japanese dialogue system by incorporating the daily life dataset PERSONA-CHAT [100] into training.

Our main contributions are summarized as follows:

- We propose a DVG model to efficiently learn the bidirectional relationship between the context and the response in the conversation for contextual understanding and affective expression. Automatic and human evaluations on both Japanese and English EMPATHETICDIALOGUES datasets show that our method outperforms competitive baselines.
- We introduce a retrieval system as a fallback to the generation process to directly produce an empathetic response. Automatic and human evaluations on the Japanese EMPATHETICDIALOGUES dataset demonstrate that compared with the solely generative model, our generative+retrieval system can generate empathetic responses with more diversity and better scores on the aspects of *Empathy*, *Relevance*, and *Fluency*.
- We evaluate our method in general response generation, which is not specific to empathetic but also chitchatting dialogue system. We further integrate our system into a virtual agent Gene, and objective automatic evaluation and subjective evaluation via human-agent interaction experiments further demonstrate our system's effectiveness.

3.2 Related Work

3.2.1 Dual Learning

Dual learning has been applied to several tasks due to its potential in improving the performance of both the primary task and auxiliary task. Tseng et al. [105] coupled natural language understanding and natural language generation through a shared latent variable, which benefits both tasks. Cui et al. [106] utilized the additional information from a response to query generation to avoid safe response. Hu et al. [107] integrated bidirectional learning with a discriminator for neural topic modeling.

In this study, we extend dual learning to efficiently learn the bidirectional relationship between context and response.

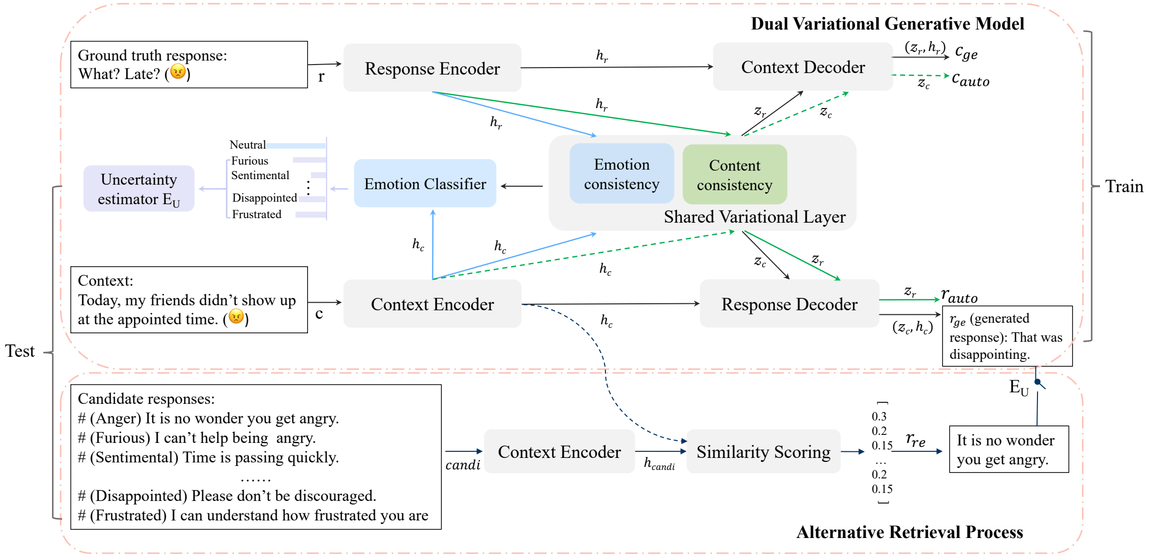


Fig. 3.2 Proposed DVG model for empathetic response generation. Blue and green highlighted lines and blocks denote the emotion consensus and content consistency processes respectively. Specifically, green solid lines represent the response reconstruction process, while green dotted lines represent context reconstruction. c_{ge}, c_{auto}, r_{auto}, r_{ge} mean context generation, context auto-reconstruction, response auto-reconstruction, and response generation, respectively. Compared with previous studies, we incorporate variational decoder to the dual generative model for context and response reconstruction (c_{auto} and r_{auto}).

3.2.2 Retrieval-based Response Generation

Retrieval-based methods have been considered as an alternative or complement to enhance the generation-based approaches. Cai et al. [108] explored a retrieval-guided response generation based on a matching mechanism. Zhang et al. [109] proposed to attentively combine retrieval and generation using a Mixture-of-Experts ensemble to generate a follow-on text. The above studies combined a retrieval system trained with a generation model, thus the effectiveness is very sensitive to the retrieval quality, which may even worsen the generation process. To avoid this problem, we adopt the retrieval system as a fallback to the generation model based on emotion classification to alleviate the difficulty of empathetic response generation.

3.3 Dual Variational Generative (DVG) Model

As shown in the green solid and dotted flows in Fig. 3.2, we incorporate a variational model into the dual generation framework. The variational decoder is utilized for not only generation but also reconstruction between the context and response, and the reconstruction process ensures the consistency and makes the learning of the shared layer easy. For a given context c, the goal is to generate an empathetic response r_{ge} by the proposed DVG. We will explain each module in the following.

3.3.1 Baseline Dual Generative Model

Our proposed DVG model is based on a dual generative model, which coupled the response generation from context and context generation from response with one duality layer. The duality layer models the mutual relationships between the context and response, such as emotion consensus [104]. The basic unit of generation module can be chosen from GRU [106], LSTM [105, 110, 111] and Transformer [104]. In this work, we utilize Transformer [47] encoder and decoder, which have shown effectiveness in various tasks [112–116].

Shen et al. [104] tried to ensure the emotion consistency from duality complementarity (with the blue circle in Fig. 3.2) and composed the shared layer with a simple dense and softmax networks. However, the variables of the dense layer are deterministic. In this paper, we design the shared layer to be variational, which allows for composing random variables to generate diverse responses. We also incorporate a variational model into the dual generative model with a reconstruction process (e.g. context to context) to enhance the shared layer for better content consistency, in addition to the consistency between the context and response.

3.3.2 DVG Model Architecture

There are two similar processes in our DVG model. One is the forward dialogue process from context to response (r_{ge}): context encoder, shared variational layer, and response decoder. The other is the backward dialogue process from response to context (c_{ge}): response encoder, shared variational layer, and context decoder. Moreover, we incorporate the variational auto-reconstruction process from context to context (c_{auto}) and from response to response (r_{auto}). We utilize Transformer for the encoders and decoders. An emotion classifier is augmented to this model. We describe the details of the forward dialogue process in this subsection.

Context Encoder

Inspired by Devlin et al. [48], we firstly add a special token `[CLS]` to the beginning of the context c, which represents the global memory of the whole sequence. Then the input context c are converted to word embeddings $\text{emb}_w(c)$, summed with the position embeddings $\text{emb}_{pos}(c)$:

$$e_c = \text{emb}_w(c) + \text{emb}_{pos}(c). \tag{3.1}$$

Finally, we employ a Transformer encoder to get the context representation:

$$h_c = \mathrm{trs}_{\mathrm{enc}_c}(e_c), \tag{3.2}$$

where $\mathrm{trs}_{\mathrm{enc}_c}$ is the forward context encoder, $h_c \in \mathbb{R}^{n \times d_{\mathrm{enc}_c}}$, n is the number of encoder layers, and d_{enc_c} is the dimension of the encoder layer.

Shared Variational Layer

We assume that there exists a continuous latent representation z, which represents the mutual characteristics, underlying a pair of context c and response r, where z can be inferred from either c or r. Considering the intractable posterior distribution of unobserved variable z, inspired by Kingma et al. [117], we choose the approximated posterior distribution $q_{\mathrm{trs}_{\mathrm{enc}_c}}(z|h_c)$ to be Gaussian, $\mathrm{trs}_{\mathrm{enc}_c}$ is the forward context encoder, and utilizes the reparameterization trick:

$$\begin{aligned} z_c &= \mu_c + \sigma_c \odot \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, I). \end{aligned} \tag{3.3}$$

Then we use the hidden layer of the context encoder output h_c to compute the variable μ and σ in the variational process:

$$\begin{aligned} \mu_c &= \omega_1 h_c + b_1, \\ \sigma_c^2 &= \omega_2 h_c + b_2, \end{aligned} \tag{3.4}$$

where ω_1, ω_2, b_1, and b_2 represent feedforward network weights and biases. For the backward response model, we do the same process:

$$\begin{aligned} z_r &= \mu_r + \sigma_r \odot \varepsilon, \\ \mu_r &= \omega_3 h_r + b_3, \\ \sigma_r^2 &= \omega_4 h_r + b_4, \end{aligned} \tag{3.5}$$

where h_r is the output of response encoder $\mathrm{trs}_{\mathrm{enc}_r}$ in the backward response model, ω_3, ω_4, b_3, and b_4 represent backward network weights and biases.

Response Decoder

We incorporate the mutual representation z into the Transformer decoder [47] for output generation. First, we add a special token SOS to the beginning of the decoder input $y_{<t}^{(i)}$, and

conduct word and positional embeddings:

$$e_t^{(i)} = \mathrm{emb}_w(y_{<t}^{(i)}) + \mathrm{emb}_{pos}(y_{<t}^{(i)}), \tag{3.6}$$

where i is in a value index of $\{r, c\}$. To efficiently learn the representation of z_c, used for response generation, we also introduce a task of context reconstruction, which involves contextual understanding, inspired by VAE [117].

$$\begin{aligned} H_{r_{ge}} &= \mathrm{trs}_{\mathrm{dec}_r}([e_t^{(r)}; h_c; z_c]), \\ H_{c_{auto}} &= \mathrm{trs}_{\mathrm{dec}_c}([e_t^{(c)}; z_c]). \end{aligned} \tag{3.7}$$

Here $\mathrm{trs}_{\mathrm{dec}_r}$ and $\mathrm{trs}_{\mathrm{dec}_c}$ correspond to the forward response decoder and the backward context decoder, respectively. $H_{r_{ge}}$ and $H_{c_{auto}}$ represent the hidden states of the response decoder and context decoder outputs, respectively. $[e_t^{(c)}; z_c]$ denotes the concatenation of $e_t^{(c)}$ and z_c. Then, we compute the generic vocabulary token distribution:

$$p(y_t^{(i)}|H) = \mathrm{softmax}(W_v H + b_v), \tag{3.8}$$

where H corresponds to the $H_{r_{ge}}$ or $H_{c_{auto}}$, and $p(y_t^{(i)}|H)$ is the output token distribution at time step t. W_v and b_v represent the weights and bias of the corresponding softmax network.

Emotion Classifier

We introduce an emotion classifier to explicitly detect the emotion from the user utterance. It is trained from the response as well. The emotion classifier is connected with the shared variational layer to achieve emotion consistency between the context and response. We use the `[CLS]` embedding h_{c_0} of the encoder output to represent the global memory of the entire context. And we use the cross-entropy as the loss function:

$$\begin{aligned} p_e &= \mathrm{softmax}(W_e[h_{c_0}; z_c] + b_e), \\ \mathscr{L}e &= \sum_{i=1}^{n_e} -e_s * \log(p_e), \end{aligned} \tag{3.9}$$

where W_e and b_e represent the weights and bias of the emotion classifier network; e_s is the ground-truth emotion label, n_e is the number of emotion categories.

3.3.3 DVG Model Optimisation

We describe how to optimize our proposed model in this sub-section. Given the paired datapoint (c,r), the main objective is to optimize the log-likelihood of the joint generation probability $p(c,r)$:

$$\mathscr{L} = \log \int p(c,r,z_c)dz_c + \log \int p(r,c,z_r)dz_r. \tag{3.10}$$

However, this optimization is intractable because of the unknown latent variable z_c and z_r. Inspired by the derivations from Tseng et al. [105], Shen et al. [104], we follow the neural variational inference as introduced in the variational Bayes approach [117]:

$$\begin{aligned}
\log \int p(c,r,z_c)dz_c =& \log \int \frac{p(c,r,z_c)q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)}{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)}dz_c \\
=& \log \int \frac{p(c|z_c)p(r|z_c,c)p(z_c)q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)}{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)}dz_c \\
=& \log \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)} \frac{p(c|z_c)p(r|z_c,c)p(z_c)}{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)} \\
\geqslant& \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)} \log \frac{p(c|z_c)p(r|z_c,c)p(z_c)}{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)} \\
=& \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)}(\log p(c|z_c) + \log p(r|z_c,c)] \\
& - D_{KL}[q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)||p(z_c)],
\end{aligned} \tag{3.11}$$

$$\begin{aligned}
\log \int p(r,c,z_r)dz_r =& \log \int \frac{p(r,c,z_r)q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)}{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)}dz_r \\
=& \log \int \frac{p(r|z_r)p(c|z_r,r)p(z_r)q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)}{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)}dz_r \\
=& \log \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)} \frac{p(r|z_r)p(c|z_r,r)p(z_r)}{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)} \\
\geqslant& \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)} \log \frac{p(r|z_r)p(c|z_r,r)p(z_r)}{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)} \\
=& \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)}(\log p(r|z_r) + \log p(c|z_r,r)] \\
& - D_{KL}[q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)||p(z_r)].
\end{aligned} \tag{3.12}$$

Then our objective can be achieved by maximizing the variational lower bound of $\mathscr{L}c,r$ and $\mathscr{L}r,c$:

$$\mathscr{L} \geq \mathscr{L}c,r + \mathscr{L}r,c, \tag{3.13}$$

where $\mathscr{L}c,r$ and $\mathscr{L}r,c$ are the objective function of the forward context model and the backward response model, separately. The former is formulated as:

$$\begin{aligned}\mathscr{L}c,r =& \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)} \log p(r|z_c,c)\\ &+ \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)} \log p(c|z_c)\\ &- D_{KL}[q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)||p(z_c)].\end{aligned} \tag{3.14}$$

The first term represents response generation in the forward process; the second term denotes the variational auto-reconstruction of context; the third term means the Kullback-Leibler (KL) divergence between the forward Gaussian posterior $q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)$ with the prior distribution $p(z_c)$ of the shared variational layer, where $q_{\mathrm{trs}_{\mathrm{enc}_c}}(z_c|c)$ and $p(z_c)$ are both the multi-variate standard Gaussian distributions. Similarly, we can derive a variational optimization objective for the backward response model:

$$\begin{aligned}\mathscr{L}r,c =& \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)} \log p(c|z_r,r)\\ &+ \mathbb{E}_{q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)} \log p(r|z_r)\\ &- D_{KL}[q_{\mathrm{trs}_{\mathrm{enc}_r}}(z_r|r)||p(z_r)].\end{aligned} \tag{3.15}$$

Finally, the entire model is optimized with the sum of $\mathscr{L}c,r$, $\mathscr{L}r,c$ and $\mathscr{L}e$.

3.3.4 Alternative Retrieval

To alleviate the difficulty of generating appropriate empathetic responses, we incorporate the retrieval process in the testing to serve as a fallback of the generation process as shown in Fig. 3.2. We first compute the emotion distributions of the input context as shown in Equation (3.9).

Then, we select the corresponding n candidate responses from the pre-defined set based on the predicted emotions, which are taken from the top five candidates of the classification probabilities. We use the same context encoder to encode the selected candidate responses:

$$h_{candi_i} = \mathrm{trs}_{\mathrm{enc}_c}(\mathrm{candi}_i), \tag{3.16}$$

where candi_i is the i-th selected candidate response, and i ranges from (1 to 5)$\times n$. Then, we compute the similarity score $\mathrm{sim}_{i,j}$ between the candidate representation h_{candi_i} and input context h_j:

$$\mathrm{sim}_{i,j} = 1 - \arccos(\frac{h_j^\top h_{candi_i}}{||h_j||\,||h_{candi_i}||})/\pi. \tag{3.17}$$

Then, candidate r_{re} is chosen by the ranking of the similarity score.

3.3.5 Uncertainty Estimator

To select a response from generation or retrieval, we estimate the emotion uncertainty of our DVG model, which is computed by the entropy of the emotion classification probabilities:

$$E_U = \sum_{v=1}^{V} p_e^v \log p_e^v, \tag{3.18}$$

where V is the number of the emotion categories. After obtaining the generated response r_{ge} and retrieved response r_{re}, we choose the best one based on a threshold u:

$$r = \begin{cases} r_{re}, & \text{if } E_U < u \\ r_{ge}, & \text{if } E_U \geq u. \end{cases} \tag{3.19}$$

3.4 Experiments on Empathetic Response Generation

We conducted evaluations in both Japanese and English EMPATHETICDIALOGUES datasets. Japanese has been selected to facilitate the implementation of our system in Japanese-speaking agents/robots, thereby facilitating the evaluation of human-agent interactions. Additionally, we opt for English due to its widespread use within this task, enabling direct comparisons with previous studies. For the retrieval process, a Japanese speaker created two or three candidate responses for each emotion category that do not depend on the context and can be used in many situations. In total, there are 82 candidate responses.

3.4.1 Settings

We set the batch size to 16 and the learning rate to 0.0001. We used JUMAN++ for Japanese word segmentation. We used pre-trained fastText [118] vectors to initialize the word embeddings. All hyper-parameters of the Transformer model were set the same as in previous work [84]. Following Shen e al. [104] and Tseng et al. [105], we applied KL annealing [119] to alleviate the degeneration issue of the variational network. We used greedy search during inference in the generation process and the maximum decoding step was set to 30.

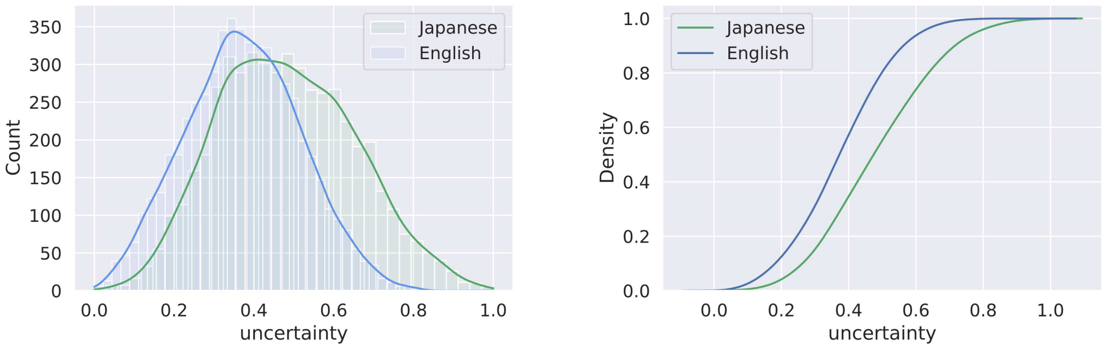
3.4.2 Comparison Models

For a comprehensive evaluation, we compare our model with other state-of-the-art models.
Transformer [60]: This is a standard Transformer encoder-decoder architecture model. After encoder, it coupled a response decoder and emotion classification.
MoEL [73]: This is an extension of Transformer, which softly combines multiple emotion-specific decoders to a meta decoder to generate an empathetic response.
MIME [74]: This method assumes that empathetic responses often mimic the speaker's emotion and integrates emotion grouping, emotion mimicry, and stochasticity into the emotion mixture for various empathetic responses.
Dual-Emp [104]: This method introduced the dual learning framework, which simultaneously constructs the emotion consensus by a dual-generative model, and also utilizes some external unpaired data. Note that, for a fair comparison, we only compare with this method without using external unpaired data. The major difference from our model is that we also incorporate a variational model into the dual generative model, using a reconstruction loss for both contexts and responses. Our model enhances the shared layer for better content consistency, in addition to the consistency between the context and response.

3.4.3 Evaluation Measures

Automatic Metrics

For automatic evaluation, we use the following metrics: (1) PPL (Perplexity) [120] measures how well a language models predicts a response, with lower values indicating better performance. (2) BLEU [121] which evaluates the matching of the generated response to the ground truth. We use *multi-bleu.perl* [122] to compute the BLEU scores. (3) EA (Emotion accuracy), which evaluates whether the model correctly recognizes emotion states. There are some similar emotions in the 32 categories. Thus, if the ground truth emotion falls into the top 5 predicted emotions, then we regard the correct prediction. (4) D1/D2 (Distinct-1/ Distinct-2) [123] to evaluate the diversity aspect. (5) BERTScore [124] is a BERT-based evaluation measure for text generation, which focus on lexical semantic similarity between the generated response and the ground truth.



(a) Count distribution (b) Cumulative distribution

Fig. 3.3 Emotion uncertainty distribution on the validation set of the Japanese and English EMPATHETICDIALOGUES datasets.

Human Evaluation

We randomly sampled 100 dialogues and their corresponding responses generated from our method as well as the compared methods. We recruited crowd-workers to evaluate the responses generated by various models. Annotators were asked to evaluate the quality of the generated response based on three dimensions: Empathy, Relevance, and Fluency [104, 74, 60]. Three crowd-workers evaluated each dimension, and we used the average value. Empathy measures whether the generated response contains the emotion understanding of the context. Relevance considers the topic consistency between the context and the generated response. Fluency assesses whether the generated responses are linguistically correct and readable. Each metric is rated on a scale from 1 to 5.

Human A/B Test

To directly compare the overall performance of our method and others, we also adopt the human A/B test. For two generated responses, one is by our DVG, and the other is from one of the compared models: Transformer, MOEL, MIME, Dual-Emp. Three annotators were asked to choose the better one, or select ‘Tie.’

3.4.4 Emotion Uncertainty Threshold

It is important to find a suitable threshold for the emotion uncertainty estimator to select the final output from the generated and retrieved responses. Fig. 3.3 depicts the count and cumulative distributions of the emotion uncertainty in the validation set. For example, we can see from the cumulative distribution that there is about 18% percent of the samples with

Table 3.1 Results of the proposed method with different uncertainty thresholds on the validation set of the Japanese and English EmpatheticDialogues datasets.

	Uncertainty threshold	Cumulative	Dist-1(%)	Dist-2(%)
Japanese	0.2	0.05	2.08	8.01
	0.3	0.18	2.21	**8.31**
	0.4	0.35	**2.29**	**8.25**
	0.5	0.50	2.28	8.06
English	0.18	0.1	2.63	8.84
	0.25	0.2	**2.66**	**8.89**
	0.30	0.3	2.67	8.77

Table 3.2 Automatic and human evaluation results of our method and compared models for the Japanese EmpatheticDialogues dataset, bold font denotes the best performances. BERT represents BERTScore. Emp, Rel, Flu are abbreviations of Empathy, Relevance, and Fluency, respectively.

Model	Automatic Evaluation						Human Evaluation		
	PPL ↓	BLEU	EA(%)	D1(%)	D2(%)	BERT(%)	Emp	Rel	Flu
Transformer [60]	20.33	**6.92**	69.25	1.34	5.77	73.21	2.88	2.47	2.89
MoEL [73]	19.49	0.66	68.69	1.36	5.67	73.36	3.15	2.74	2.95
MIME [74]	20.69	0.64	62.46	0.69	2.62	73.08	3.22	2.77	3.24
Dual-Emp [104]	19.23	6.91	71.89	1.11	3.66	73.29	3.22	2.89	**3.30**
DVG (Ours)	**18.32**	6.79	**74.29**	**2.06**	**7.94**	**73.57**	**3.47**	**3.22**	3.24

emotion uncertainty smaller than 0.3, which means if the emotion uncertainty threshold is set to 0.3, 18% percent of generated responses will be replaced by the corresponding retrieved one. Based on the values of *D1* and *D2* in Table 3.1, we chose the emotion uncertainty threshold to be 0.3 or 0.4 for the Japanese experiments and 0.25 for the English experiments.

3.4.5 Japanese Dialogue Results and Analysis

Comparison with other Methods

The automatic evaluation results in the left part of Table 3.2 show that our DVG model outperforms others in the aspects of *emotion accuracy (EA)*, and diversity metrics (*D1* and *D2*). It demonstrates our model's potential to detect emotions more effectively considering both the emotion and content consistency between the context and response, as well as the ability to generate more diverse responses.

Table 3.3 Results of human A/B test for the Japanese EMPATHETICDIALOGUES dataset.

DVG (ours) vs.	Win	Loss	Tie
Transformer	42.7%	21.7%	35.7%
MoEL	38.3%	28.7%	33.0%
MIME	38.3%	29.7%	32.0%
Dual-Emp	35.0%	29.0%	36.0%

Table 3.4 Evaluation of the alternative retrieval system for the test set of Japanese EMPATHETICDIALGUES dataset.

Model	Automatic Evaluation			Human Evaluation		
	BLEU	Dist-1 (%)	Dist-2 (%)	Empathy	Relevance	Fluency
DVG	**6.79**	2.06	7.94	3.47	3.22	3.24
DVG + Retrieval (E_u=0.3)	5.74	2.18	**8.14**	**3.67**	**3.41**	3.71
DVG + Retrieval (E_u=0.4)	6.23	**2.20**	7.99	3.50	3.29	**3.96**

Human evaluation results in Table 3.2 indicate that, among the compared models, our DVG model has the best performance with more than 7.76% and 11.42% improvement on the dimensions of *Empathy* and *Relevance*, respectively. It confirms our model's superiority for suitable emotion and content expression. Especially, compared with *Dual-Emp*, the improvement on the *Relevance* aspect is noteworthy, which indicates that our model can generate responses with contextual appropriateness.

In addition, we conducted pairwise comparisons between *DVG* with the baseline models to directly compare the overall quality of the generated responses. The results of the human A/B test in Table 3.3 show that the proposed *DVG* is significantly preferred over others by human judges.

Effectiveness of the Alternative Retrieval System

The effectiveness of the alternative retrieval process is shown in Table 3.4 using the test set. Compared with the generative model *DVG*, both *DVG + Retrieval (*E_u*=0.3)* and *DVG + Retrieval (*E_u*=0.4)* are superior in the automatic evaluation metrics of *Dist-1*, *Dist-2* and human evaluation metircs of *Empathy*, *Relevance*, and *Fluency*. The *BLEU* score is degraded, because there exists a discrepancy between the retrieved statement and the ground-truth response, but the result demonstrates our model generates relevant empathetic responses with better diversity and higher scores by human evaluation. It confirms the effectiveness of the

Table 3.5 Sample responses generated by our proposed *DVG+ Retrieval* and other models (Original in Japanese and translated to English).

Ground truth emotion	誇りに思う Proud
Predicted emotion	‘誇りに思う’, ‘感謝する’, ‘感動する’, ‘羨ましい’, ‘感傷的になる’ ‘proud’, ‘grateful’, ‘impressed’, ‘jealous’, ‘sentimental’
Context	SPK1:大学でお世話になった教授の論文が著名な雑誌で発表になったよ。 A paper by one of my professors in college has just been published in a prominent journal. SPK2: それは嬉しいね! That’s nice to hear. SPK1:うん。もう今は何の関係もないけど誇りに思うわ。 Yeah, we don’t have a relationship anymore, but i am proud of it.
Ground truth response	本当だね。嬉しいし誇らしいね。 It’s true, happy and proud.
Transformer	すごいね、私もそういうこともあるよね。 That’s great, I know that happens to me too.
MoEL	すごいね、頑張って。 That’s great. Good luck.
MIME	そうなんだね。私も一緒に見てみたいな。 I see. I’d like to see it with you.
Dual-Emp	そうだね。 Yeah, right.
DVG (Ours)	そうだね、 **応援しているんだね。** Yeah, right. **You’re cheering them on**.
Retrieval (Ours)	すばらしいですね。 That’s wonderful.
DVG+Retrieval (Ours)	すばらしいですね。 That’s wonderful.
Ground truth emotion	うしろめたい Guilty
Predicted emotion	‘さびしい’, ‘うしろめたい’, ‘楽しい’, ‘満足’, ‘感傷的になる’ ‘lonely’, ‘guilty’, ‘happy’, ‘contented’, ‘sentimental’
Context	専業主婦なのに、一日ゲームにはまっちゃって何にもできなかったよ。 I’m a housewife, but I’ve been stuck in the game all day and haven’t been able to do anything.
Ground truth response	別にそういう日もあってもいいと思うよ。 I think it’s okay to have such a day.
Transformer	それはすごいね。 That’s amazing.
MoEL	それは良かったね。お子さん、何かあったの? That’s good. What happened to your child?
MIME	そうなんだ。それは楽しみだね。 I see. That’s something to look forward to.
Dual-Emp	そうなんだ。それは残念だったね。 I see. That was bad.
DVG (Ours)	そうなんだ。**でも、たまにあるよね。** I see. **But it does happen sometimes**.
Retrieval (Ours)	気まずいですね。 That’s awkward.
DVG+Retrieval (Ours)	そうなんだ。でも、たまにあるよね。 I see. But it does happen sometimes.

plug-and-play retrieval process as an alternative to the generation of the method based on the emotion uncertainty estimation.

In addition, we can see that the emotion uncertainty threshold set to 0.3 is superior to one with 0.4 in the aspects of *Empathy* and *Relevance*, inferior in *Fluency*. There is an apparent trade-off between the Empathy/Relevance and Fluency, which can be controlled by the threshold on E_u because the retried responses are always fluent but not necessarily

Table 3.6 Automatic evaluation results of our method and compared models for the English EMPATHETICDIALOGUES dataset, bold font denotes the best performances. BERT represents BERTScore.

Model	Automatic Evaluation						Human Evaluation		
	PPL ↓	BLEU	EA (%)	D1 (%)	D2 (%)	BERT (%)	Emp	Rel	Flu
Transformer [60]	37.33	2.61	73.0	2.17	7.78	85.74	3.44	3.07	3.60
MoEL [73]	37.63	2.53	68.13	1.75	6.51	85.91	3.51	3.19	3.46
MIME [74]	36.84	2.51	69.65	1.68	6.21	**85.91**	3.47	3.45	3.66
Dual-Emp [104]	34.52	**2.67**	69.82	1.38	3.96	85.89	**3.59**	3.40	3.63
DVG (ours)	**32.18**	2.61	**75.83**	**2.42**	**8.26**	85.85	3.53	**3.56**	**3.76**

Table 3.7 Results of human A/B test for the English EMPATHETICDIALOGUES dataset.

DVG (ours) vs.	Win	Loss	Tie
Transformer	47.0%	24.7%	28.3%
MoEL	45.3%	34.3%	20.3%
MIME	41.3%	36.3%	22.3%
Dual-Emp	43.3%	25.7%	31.0%

relevant and empathetic due to occasional errors in emotion recognition. This choice hinges on the intended tasks that the system is meant to accomplish. When the aim is for the system to function as a casual chit-chat bot for everyday conversations, emphasizing fluency might take precedence over showcasing emotional support (empathy). On the other hand, instances where the user necessitates distinct emotional support from the system, such as in the role of a companion for an elderly or depressive individual, would warrant a greater emphasis on the system's empathetic capabilities.

Case Studies

To illustrate the effectiveness of our proposed *DVG*, we present two examples, as shown in Table 3.5. In the first case, compared with the baselines, our proposed *DVG* generates a response of "そうだね ∼ *(Yeah, right.)*" to show cognitive understanding of the context and then "応援しているんだね。*(You are cheering them on.)*" to show the empathy as responding in the perspective of the counterpartner. As the emotion uncertainty of this sample is low, we use the retrieved response "すばらしいですね。*(That's fantastic.)*" which is matched to the predicted emotion as the final output.

In the second case, compared with *Transformer*, *MoEL*, and *MIME*, which misunderstand user's situation and emotion, *Dual-Emp* which also utilizes a dual generative model generates an empathetic expression with suitable content. Compared with *Dual-Emp*, our model which additionally utilizes VAE to mimic the process of context/response understanding by reconstruction can generate more context relevant and emotional comfort response, as "でも、たまにあるよね。*(But it does happen sometimes.)*" helps relieve the speaker's guilt. In this case, the emotion uncertainty is high, therefore, we adopt the generated response as the final response.

3.4.6 English Dialogue Results and Analysis

Automatic evaluation in Table 3.6 and human evaluation results in Table 3.7 on the English EMPATHETICDIALOGUES dataset indicate our model's superiority over the baselines.

To have an intuitive demonstration, we also list two cases in Table 3.8. Case 1 shows that *MoEL*, *MIME*, *Dual-Emp* and our *DVG* model can detect the right emotion and show emotional expression, like "*Oh no! I am sorry to hear that.*" However, our *DVG* model generates a response with better empathy, as "*I hope you are okay.*" also expresses concern. In this case, the emotion uncertainty is high, therefore, we adopt the generated response as the final response. Case 2 illustrates our model's ability for context understanding, and the response "*What happened?*" is similar to the ground truth. As the emotion uncertainty of this sample is low, we use the retrieved response *"That's tough."* which is matched to the predicted emotion as the final output.

Analyzing automatic and human evaluations of both English and Japanese (refer to Tables 3.2, 3.3, 3.6, 3.7), we can see that our approach consistently demonstrates competitive efficacy in both languages.

3.5 Experiments on General Response Generation

In daily-style conversation, an empathetic response is just one kind of conversation reply, while neutral chatting also accounts for a large percentage. Therefore, we further applied our model to a general Japanese dialogue system for the aim of human-robot interaction by incorporating the daily life dataset PERSONA-CHAT [100] into training. `PersonaChat` focuses on the personality about which the other speaker talks without emotion.

3.5.1 Settings

We trained and evaluated the model for each turn of *Listener* responding to *Speaker*, and extended *Speaker*'s inquiries one by one from the context history. We trained the model with 33 emotion categories, which consist of 32 emotions used in the EMPATHETICDIALOGUES

Table 3.8 Case studies of the proposed DVG model and other models on the English EmpatheticDialogues dataset.

Ground truth emotion	Anger
Predict emotion	Apprehensive, Embarrassed, Ashamed, Afraid, Guilty
Context	I just moved to this neighborhood and some dumb criminals shot one of my neighbors and ran into the woods!
Ground truth response	That's not good. Do you own a gun?
Transformer	Oh my goodness! i would be too.
MoEL	Oh no! That is so annoying! I am glad you are able to do it!
MIME	Oh no! That is terrible. I would have been so scared.
Dual-Emp	Oh no! I am sorry to hear that. I am sorry to hear that.
DVG (Ours)	Oh no! I am sorry to hear that. **I hope you are okay**.
Retrieval (Ours)	I hope everything will be fine.
DVG+Retrieval (Ours)	Oh no! I am sorry to hear that. I hope you are okay.
Ground truth emotion	Sad
Predict emotion	Sad, Devastated, Lonely, Disappointed, Faithful
Context	I am so sad because I have read an article about a newborn girl who died because her parents did not believe in medication and doctors.
Ground truth response	ugh, those articles always get me too ... : (what was wrong with her?
Transformer	Oh no! I am so sorry to hear that.
MoEL	I am so sorry to hear that. I hope you are able to find her.
MIME	That is sad. I am so sorry to hear that.
Dual-Emp	That is terrible, I am so sorry to hear that.
DVG (Ours)	I am so sorry to hear that. **What happened?**
Retrieval (Ours)	That's tough.
DVG+Retrieval (Ours)	That's tough.

dataset and one additional neutral category for experiments. All the experimental settings were the same as those described in Section 3.4.1.

3.5.2 Objective Evaluation

The results in Table 3.9 show that our model trained with the two datasets does not degrade for each of them. It means our model can generate both empathetic and neutral responses for a general dialogue system. In fact, combining the two datasets contributes to an overall improvement over using a single dataset, even though the topics and emotions are significantly different.

Table 3.9 Objective evaluation results when combining PERSONA-CHAT ('Persona') with EMPATHETICDIALOGUES ('Empa') dataset.

Training	Testset	PPL	BLEU	EA (%)	D1 (%)	D2 (%)	BERT (%)
Empa	Empa	10.97	23.04	70.33	2.08	6.75	79.37
Persona	Persona	16.74	16.38	-	2.36	8.56	73.48
Empa+Persona	Empa+Persona	12.62	19.6	-	2.95	9.42	77.29
	Empa	**10.11**	**24.92**	68.21	**2.43**	**7.97**	79.26
	Persona	**15.44**	16.24	-	**2.92**	**9.37**	**75.50**

3.6 Implementation

3.6.1 Reference: Attentive Listening System

For reference, we compared with an attentive listening system [6], which can generate several types of listener responses: backchannels, repeats, elaborating questions, assessments as well as empathetic responses. The system is reported to show comparable performance to the WOZ system in basic skills of attentive listening such as actively listening, encouragement to talk, and focused on the talk.

3.6.2 Empathetic Attentive Listening System

We further implemented our retrieval model to the previous attentive listening system [6] for the autonomous android ERICA [7] to make an empathetic attentive listening system. For each input utterance, we recognized its emotion and computed emotion uncertainty; if it was lower than the predefined threshold, we selected one response by the retrieval system; otherwise used the responses like backchannels, partial repeats as defined in Inoue et al. [6] to continue the conversation.

Case Studies

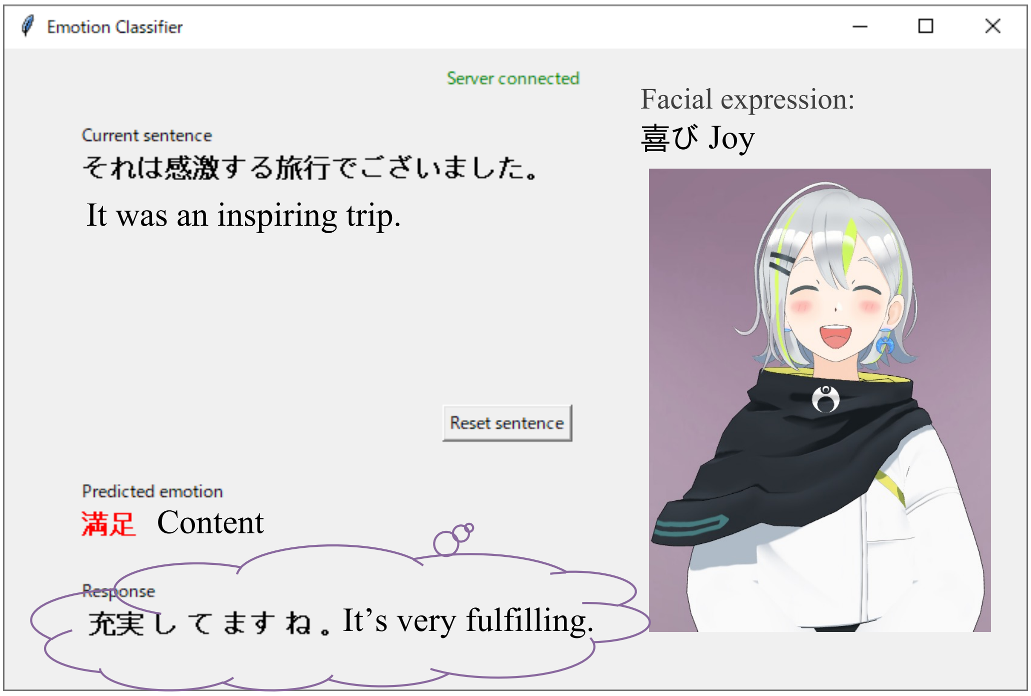
We present a comparation between our retrieval system and sentimental responses in the previous attentive listening system in Table 3.10. We can see that compared with the previous system, which produces four kinds of sentimental responses based on positive and negative sentiments, our retrieval system can generate fine-grained empathetic responses corresponding to different detected emotions. Therefore, our model can improve human-robot interaction experiences.

Table 3.10 Case studies of incorporating our retrieval system with the attentive listening system to achieve a new empathetic attentive listening system.

Input	Baseline	Retrieval system	Predicted emotion
後々不安ということでね。 I mean, I'm worried about it later on.	大変ですね。 It's hard.	不安ですね。 I'm worried about it.	不安に思う anxious
怖いからなんかこうした時に玄関の靴を持って逃げ出したりしたことがあったんです。 I used to run away with the shoes on the front porch when I was scared of something like this.	大変ですね。 It's hard.	恐怖を感じますね。 I feel scared.	怖い terrified
あまりお付き合いできないのは,もう大変残念で悲しいこと です。 It's a great pity and a great sadness that we can't socialise more.	大変ですね。 It's hard.	悲しいですね。 That's sad.	悲しい sad
一緒に話したりするのとはちょっと感覚が違ったから、多少孤独を感じた場面あったかな。 It felt a bit different from talking together, so there were times when I felt a bit lonely.	残念でしたね。 That was bad.	センチメンタルですね。 That's sentimental.	感傷的になる sentimental
やっぱ どうしても、みんなからちょっと嫌な目で見られたりと か。 I can't help but get a bit of a disapproving look from people.	残念でしたね。 That was bad.	それは嫌な気持ちになりますね。 That's a disgusting feeling.	嫌悪感を抱く disgusted
雑談というか言うて元気な姿を見せるのが一番。 The best way is to chat with them, or rather, say something and show them how healthy you are.	素敵ですね。 It's nice.	気分いいですね。 Feels great.	楽しい joyful
それは感激する旅行でございました。 It was an inspiring trip.	素敵ですね。 it's nice.	充実してますね。 It's very fulfilling.	満足 content
外国旅行いろいろ楽しい思い出はいろいろ思い出してきているんで I have many pleasant memories of traveling abroad.	いいですね。 It's good.	楽しそうですね。 Sounds like a lot of fun.	楽しい joyful
ほんとに毎週土曜日が楽しみだったんです。 I really looked forward to it every Saturday.	いいですね。 It's good.	待ち遠しいですね。 I can't wait for it.	わくわくする excited
あんまり楽しみくなりましてね。 I'm really looking forward to it.	いいですね。 it's good.	楽しみですね。 I'm looking forward to it.	期待する anticipating

Combination with Multi-modal Facial Expression

In a human-robot interaction system for multi-modal expressions, such as audio, facial, and motion, can significantly improve user experiences. We further combined our model with the facial expression for a virtual agent Gene [103] to produce vivid empathy. We can see from Fig. 3.4 that for the input utterance "それは感激する旅行でございました。 (It was an inspiring trip)", the system has an emotion prediction as "満足 (content)" and its emotion



Current sentence
それは感激する旅行でございました。
It was an inspiring trip.

Predicted emotion
満足 Content

Response
充実してますね。It's very fulfilling.

Facial expression:
喜び Joy

Fig. 3.4 Example of response with multi-modal facial expression of the virtual agent Gene.

uncertainty is lower than the selection threshold, so the system outputs retrieved response as "充実してますね。(It's very fulfilling.)", as well as a "喜び (joy)" facial expression.

3.6.3 Open-domain Chatting System

Finally, we evaluated the system in real interaction with human subjects via speech.

System Settings

For each conversational input, if the detected emotion is *neutral*, we adopt a generated response rather than using the retrieval system. Otherwise, as described in Equation (3.19), if the uncertainty of the detected emotion is smaller than the threshold E_u, we apply the retrieval system.

The context history is important for the system to understand the subject's talking, then generate a consistent and coherent response throughout the conversation. However, utterances observed by the spoken dialogue system are different from those during the model training which uses clear texts, and it is affected by the errors of the ASR system. Therefore, we make two settings of our *DVG+Retrieval* for comparison: one sets context history to 1 and the other sets to 2.

Table 3.11 Average scores on subjective evaluation and t-test results (subjects = 21). 'D+R' and 'A' represent our 'DVG+Retrieval' system and 'Attentive Listening' system, respectively. p_1, p_2, and p_{12} mean p-value of the comparison between 'A' and 'D+R (context=1)', 'A' and 'D+R (context=2)', and 'D+R (context=1)' and 'D+R (context=2)', respectively.

Metric name	Questionnaire Items	A	Context=1		Context=2		
			D+R	p_1	D+R	p_2	p_{12}
Humanness	The system's utterances were human-like and natural.	4.0	4.1	.858	3.8	.658	.498
Cognition	The system understood the talk.	4.0	3.8	.186	3.8	.229	1.00
Emotion	I felt that the system can express various emotions.	4.3	4.2	.800	4.1	.470	.479
Empathy	The system was able to empathize with my experiences.	4.7	4.2	.107	4.0	.017	.217
Personality	I felt that the system has personality.	3.7	3.9	.530	3.6	.812	.322
Agency	I felt that the system was speaking from its own perspective.	2.6	3.6	$.002^{**}$	3.3	$.031^{*}$	.507
Topic	I felt that the system had a topic it wanted to discuss.	2.1	3.2	$.001^{**}$	2.9	$.020^{*}$	.464
Attentiveness	The system was attentive to me and was actively trying to talk with me.	3.0	4.1	$.037^{*}$	3.8	$.049^{*}$	.685
Diversity	The system was able to provide various responses.	3.6	4.3	$.061^{+}$	4.2	$.085^{+}$	.893
Engagement	I felt absorbed in the interaction with the system.	3.0	3.7	$.079^{+}$	3.1	.642	.126
Ease	It was easy to continue a conversation with the system.	2.9	3.5	$.094^{+}$	3.1	.448	.185
Enjoyability	I enjoyed speaking with the system.	3.3	3.6	.425	3.1	.463	$.046^{*}$
Talk again	I want to talk with the system again.	3.1	3.4	.464	3.0	.825	.249

$(**p < .01, *p < .05, +p < .1)$

Experiment Settings

We recruited 21 students from our university for the human-agent experiment, and each subject was given the topic of *"The experience that impressed you most or recently."* but not constrained to this topic. They were asked to talk with the *Attentive Listening System* [6] and our *DVG+Retrieval* (context=1) and *DVG+Retrieval* (context=2) systems, alternately, based on the given topic. Either system was integrated into the virtual agent Gene [103]. And each conversation lasted 8 minutes. After the conversion, each subject completed the questionnaire on a point ranging from 1 (completely disagree) to 7 (completely agree) for each item, as shown in Table 3.11. The order of the test system was randomized for each subject.

Results of Human Interactions

Table 3.11 reports the average score for each question item. The *DVG+Retrieval* system performs overall better than the *Attentive Listening* system. It performs better when the context history is set to 1 than when it is set to 2, but the p_{12} value shows that there is no significant difference between the *DVG+Retrieval* (context=1) and *DVG+Retrieval* (context=2) systems. We observed subjects often switch emotions or topics within the conversation, in this case, the *DVG+Retrieval* (context=2) system tends to generate inappropriate responses because both emotion and topic are consistent within each conversation in our training datasets.

Specifically, the *DVG+Retrieval* (context=1) system achieved a significantly better score than the *Attentive Listening* system for the evaluation of *Agency*, *Topic*, *Attentiveness*, *Diversity*, *Engagement*, *Ease*. This indicates that the *DVG+Retrieval* (context=1) system can enrich the conversation with diverse topics and responses as well as be actively attentive to users. No significant difference was observed between the two systems under the evaluation of *Humanness*, *Cognition*, *Emotion*, *Empathy*, *Personality*, *Enjoyability* and *Talk again*. The *Attentive listening* system focused on keyword detection of the input, producing template-based responses. Thus, it tended to produce safe but stereotypical responses. On the other hand, the proposed system can generate more diverse responses depending on the context. But it was prone to ASR errors and often resulted in irrelevant responses.

To further examine the effect of the retrieval model in the *DVG+Retrieval* (context=1) system, we calculated the ratio between the retrieved responses against all responses. When we picked up the sessions when the retrieval ratio was larger than 10%, the *DVG+Retrieval* (context=1) system was preferred by humans over the *Attentive Listening* system in all of the subjective evaluations. This suggests that when confident emotion recognition is performed, the system works much better.

Practical Application

We incorporated our system into a CommU humanoid robot to make it practically applicable. This integration resulted in an empathetic conversational robot that is skilled in effectively engaging with humans through spoken communication.

Future Perspective

To take advantage of both *Attentive Listening* and *DVG+Retrieval*, we plan to build a hybrid system combining both systems. Specifically, we take the *Retrieval* system as the first priority to produce an emotion-specific response when the system is confident about the recognized emotion. *Attentive Listening* system is in the second priority if it generates a response in the type of ‘Repeat’ or ‘Questions’, which is safe and relevant to the context. In other cases, we can turn to the *DVG* system, which can enrich the conversation with diverse topics and responses.

3.7 Conclusions

In this paper, we have proposed the DVG model for empathetic response generation. Our DVG model can efficiently capture the mutual characteristics of the content and emotion consistency between the context and the response. Evaluations on both Japanese and English EMPATHETICDIALOGUES datasets demonstrate our model’s superiority in generating empathetic responses with contextual and emotional appropriateness. In addition to the DVG model, we proposed an auxiliary retrieval system to improve empathetic response generation. We further extended our model’s potential to generate both empathetic and general responses and evaluated our system’s effectiveness in enhancing human-robot interaction by a virtual agent. Subsequently, we integrated the system into a CommU humanoid robot for practical application.

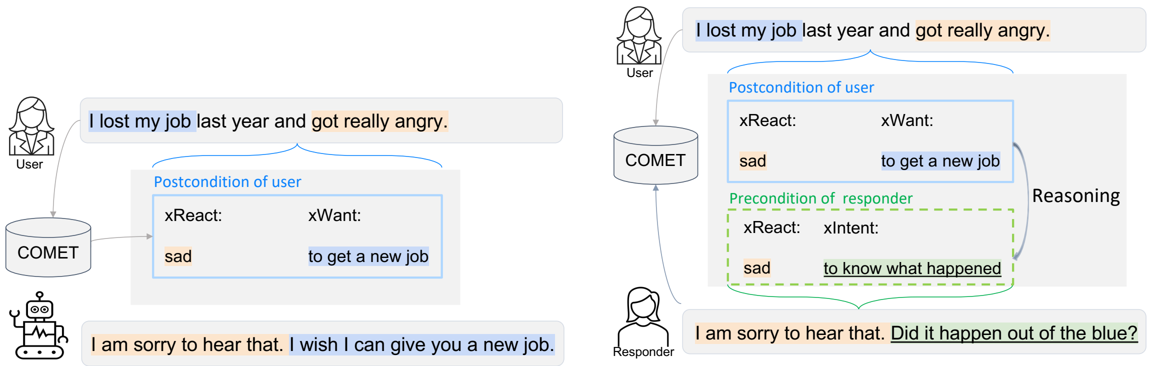
Chapter 4

Integrating Causality Reasoning For Empathetic Response Generation

4.1 Introduction

Empathy is a desirable capacity of humans to place themselves in another's position to show understanding of his/her experience and feelings and to respond appropriately. Empathy involves both cognitive and affective aspects [12], including the ability to perceive the user's situation and to express appropriate emotions.

Previous work on empathetic response generation has primarily focused on the affective aspect of emotional expression [73, 74, 125] through emotion detection, without sufficient consideration of context understanding. Recently, there has been a growing interest in exploring context understanding by leveraging external commonsense knowledge for reasoning emotion causes-effects or the user's desires, such as Sabour et al. [80] and Wang et al. [126, 77]. However, these approaches focus on understanding the causalities from the user's perspective. Exploring the causality within the user's context and reasoning his/her desires can be helpful so that the system's intention is aligned with the user's desires, and the response is generated from the user's perspective (Fig. 4.1a). However, in real human communication, the responder's intention is not always confined to the user's desires, as shown in Fig. 4.1b. Relying solely on the user's desire to generate a response may not fully understand the user's experience, and leads to weak empathy, as shown in Fig. 4.1a. Therefore, it is necessary to incorporate both the user's perspective (exploring his/her desire and reaction) and the system's perspective (reasoning its intention and reaction to mimic humans) for empathetic response generation. Through the utilization of COMET [18], which is a pre-trained GPT-2 model (Radford et al. 2018) fine-tuned on the if-then reasoning graph


User
I lost my job last year and got really angry.
COMET
Postcondition of user
xReact:
xWant:
sad
to get a new job
I am sorry to hear that. I wish I can give you a new job.
User
I lost my job last year and got really angry.
Postcondition of user
xReact:
xWant:
sad
to get a new job
COMET
Reasoning
Precondition of responder
xReact:
xIntent:
sad
to know what happened
Responder
I am sorry to hear that. Did it happen out of the blue?

(a) Example of using commonsense knowledge from COMET to generate a response from the user's perspective.

(b) Example of a response from the actual responder's perspective, based on reasoning reaction and intent to mimic humans.

Fig. 4.1 Two examples to produce a response from different perspectives. The blue solid box contains "xReact" and "xWant" representing the user's emotional reaction and desires. The green dotted box comprises "xReact" and "xIntent," representing the emotional reaction and intention of the actual responder.

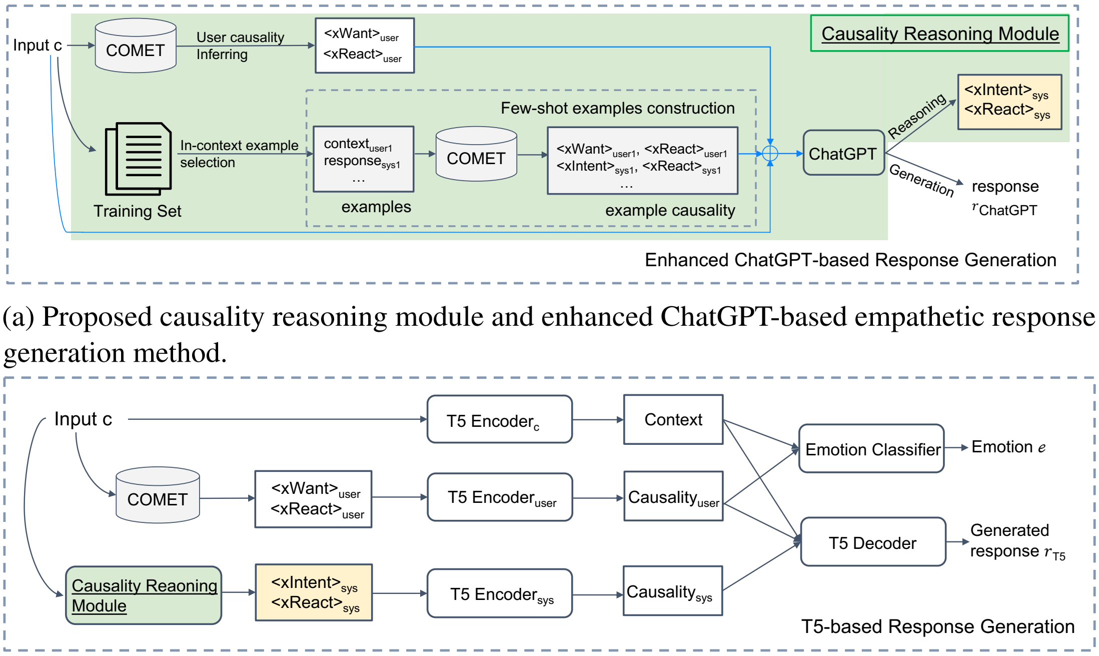
from ATOMIC [20], the system's possible intentions can be predicted to align with the user's desires. However, the system's intention may not be constrained by the user's desire. Therefore, we do not adopt COMET for the system's intention reasoning. ChatGPT[1] has shown its efficacy in several tasks [21]. Bang et al. [22] introduced ChatGPT's potential in causal reasoning on the human-annotated explainable CAusal REasoning dataset (E-CARE) [23]. However, it is based on whether the model can make a judgment on correct causes or effects instead of generating causality explanations. In this paper, we propose to enhance it by incorporating in-context learning with commonsense reasoning for causality explanation. Our main contributions are as follows:

- We propose to integrate a commonsense-based causality reasoning for empathetic response generation, which takes the system's intention and reaction, along with the user's desire and reaction.

- We propose to enhance ChatGPT's capability for causality explanation through the integration of in-context learning with commonsense knowledge (desire, reaction, and intention).

- We present experimental results to demonstrate both ChatGPT and a T5-based model, integrated with the proposed commonsense-based causality explanation, outperform other competitive methods based on both automatic and human evaluations.

[1] https://chatgpt.com/



(a) Proposed causality reasoning module and enhanced ChatGPT-based empathetic response generation method.

(b) Integrating the causality reasoning module into a T5-based encoder-decoder for empathetic response generation.

Fig. 4.2 Overview of our proposed model. The input c ends with the user's utterance. The generated response r_{T5} and $r_{ChatGPT}$ are in the role of the system (sys).

4.2 Preliminaries

4.2.1 Knowledge Acquisition

In order to generate commonsense inferences for given events, we adopt a modified BART-based [51] variation of COMET, which was trained on the ATOMIC-2020 dataset [127]. This model is suitable for inferring knowledge regarding unseen events [127], like events in the EmpatheticDialogue dataset [65].

In the training process, we leverage this model to infer the relations of *xWant* and *xReact* for each user's utterance in the training set and the relations of *xIntent* and *xReact* for the system's utterance, which are inferred from the ground-truth response in training. In the testing, we only infer the relations of *xWant* and *xReact* for the user's utterance. The system's *xIntent* and *xReact* will be inferred by the proposed causality reasoning module.

4.2.2 In-Context Example Selection

We enhance ChatGPT's causality explanation based on the few-shot setting. Given the sensitivity of large language models such as ChatGPT to in-context examples [128, 90], we adopt a method similar to Lee et al. [90] to select top-*k* examples from the training set based on the similarity between the test conversation and the training conversations. Specifically, we adopt Sentence BERT introduced by Reimers and Gurevych [129] to encode the sentence semantics of the conversation. In this study, we compute the cosine similarity between the situation utterance of the training set and the test sample, which is annotated in the dataset. Top-*k* samples are chosen from the training set for each test sample as in-context few shot examples for ChatGPT.

4.3 Proposed Method

Fig. 4.2 shows an overview of our proposed method. It consists of three components: (1) Causality reasoning module, which aims to enhance the ChatGPT or T5 decoder with a causality explanation for empathetic response generation. (2) Enhanced ChatGPT-based response generation. (3) T5-based response generation, which is based on a trained T5 encoder-decoder to be compared with other approaches that have developed their own model using the EmpatheticDialogue dataset [73, 74, 125, 80, 87].

Table 4.1 Few-shot examples (top-2 examples).

Test input	user: I'm so excited because I'm finally going to visit my parents next month! I didn't see them for 3 years.	
Few-shot1	context1	user1: Someone is visiting me soon and I can't wait! sys1: Who is it? user1: My mom, she is amazing.
	example causality	<xWant>$_{user1}$: to have a good time. to talk to their mom. to have fun with Mom. <xReact>$_{user1}$: excited. happy. satisfied. good. loved. <xIntent>$_{sys}$: to be with her. to be loved. to be nice. happy. <xReact>$_{sys}$: happy. excited. proud. good. loving.
	response1	sys1: I bet she is! I am so glad you get to see her. Mom's are awesome!
Few-shot2	context2	user2: My family is coming to visit! sys2: Awesome. When are they coming and for how long? user2: They are coming next year from Africa!
	example causality	<xWant>$_{user2}$: to have a good time. to go to the airport. to have fun with the family. <xReact>$_{user2}$: happy. excited. happy. excited. loved. <xIntent>$_{sys2}$: to see the sights. to be with family. to be with them. to have fun. <xReact>$_{sys2}$: happy. excited. satisfied. tired. relieved.
	response2	sys2: That's a long trip. I hope they have a good time.

4.3.1 Causality Reasoning Module based on ChatGPT

As outlined in Algorithm 1, this module consists of four steps. Initially, for a test input c, we employ the method outlined in Section 4.2.2 to select the top-k relevant training samples, denoted as $\mathscr{S}$, for in-context learning, such as (context1, response1) and (context2, response2) as exemplified in Table 4.1.

In the second step, for each selected sample $(c_n, r_n) \in \mathscr{S}$, we leverage the COMET model to infer the *xWant* (c_{nWant}) and *xReact* (c_{nReact}) knowledge corresponding to the user's utterance c_n. Additionally, we extract the *xIntent* ($r_{nIntent}$) and *xReact* (r_{nReact}) knowledge pertaining to the ground truth system response r_n. This information is then concatenated as few-shot examples (Table 4.1), denoted as $\mathscr{M}_{prompt}$.

Thirdly, for the test input c, we obtain the *xWant* (c_{Want}) and *xReact* (c_{React}) knowledge using COMET. Finally, they are appended to $\mathscr{M}_{prompt}$ as the prompt to ChatGPT, which reasons *Intent* (r_{Intent}) and *React* (r_{React}) from the system's perspective based on the few-shot learning.

Table 4.2 Introduction template to ChatGPT for causality reasoning and empathetic response generation.

```
Introduction:
Assuming that you are sys, who is a friend of the user. You are empathetic sometimes.
In this task, you are given the user's input and the information of "user wants to:" and "user reacts to:":
"user wants to:", which means what the user wants to do after the input;
"user reacts to:", which means how the user react to the input.

After that, please reason about the following two parts:
"sys's intent:": which means what the sys wants to do after the input, or what's the intent of sys to respond to the input;
"sys reacts to:", which means how the sys reacts to the input.

Then you respond (should be concise, no more than 30 words) to the input based on the information
of user's input, "user wants to:", "user reacts to:", "sys's intent:", "sys reacts to:".

"sys:": which means the response of sys.

Please generate the following three parts in the format below:
sys's intent:
sys reacts to:
sys:
```

4.3.2 Enhanced ChatGPT-based Response Generation

The prompt provided to ChatGPT encompasses two components: causality explanation from the user's perspective, predicted by COMET, and causality explanation from the system's perspective, derived through the causality reasoning module described in Section 4.3.1. These components, along with the introduction shown in Table 4.2 and few-shot examples, are integrated into ChatGPT to generate empathetic responses.

Algorithm 1 Commonsense-based causality explanation prompt

```
Require: A training set $\mathscr{D}=\{(c_n,r_n)\}_{n=1}^{N}$, $N$ is the number of training samples; a test input
    ($c$); $c$, $r$ represents context, ground truth response, respectively; COMET model $f_\theta(\cdot)$
  /*Step 1: In-context examples selection*/
  $\mathscr{M}_{sim} \leftarrow$ empty list
  for each $d=(c_n,r_n) \in \mathscr{D}$ do
      Get similarity score: $sim_n$
      $\mathscr{M}_{sim}$.append($sim_n$)
  end for
  $\mathscr{S}=\{(c_n,r_n)\}_{n=1}^{k}$=max($\mathscr{M}_{sim}$,$k$), $k$ is the number of in-context examples
  /*Step 2: Get the commonsense knowledge for the selected examples */
  $\mathscr{M}_{prompt} \leftarrow$ empty list
  for each $s \in \mathscr{S}$ do
      Get causality information (desire and reaction of user, intent, and reaction of sys) for
  the sample in $\mathscr{S}$ inferred by COMET
      $c_{nWant}= f_\theta(c_n + [xWant])$
      $c_{nReact}=f_\theta(c_n + [xReact])$
      $r_{nItent}=f_\theta(r_n + [xIntet])$
      $r_{nReact}=f_\theta(r_n + [xReact])$
      $k_n=c_{nWant}+c_{nReact}+r_{nIntent}+r_{nReact}$
      $\mathscr{M}_{prompt}$.append($c_n$,$k_n$,$r_n$)
  end for
  /*Step 3: Get the commonsense knowledge for the test sample */
  Get causality information (desire and reaction of user) for the test sample $c$
  $c_{Want}= f_\theta(c + [xWant])$
  $c_{React}=f_\theta(c + [xReact])$
  /*Step 4: prompting ChatGPT, and output the reasoned Intent, React for generating
  a empathetic response*/
  Input: $\mathscr{M}^{+}_{prompt}=\mathscr{M}_{prompt}+c+c_{Want}+c_{React}$
  Output: $r_{Itent}$, $r_{React}$, $r_{ChatGPT}$
```

4.3.3 T5-Based Response Generation

Context and Causality Encoding For a test input c, we use the COMET model to infer the user's causality information, which are desire and reaction of the user (k_{user}: c_{Want} and c_{React}), and use the causality reasoning module based on ChatGPT to infer the system's causality information, which are intention and reaction of the system (k_{sys}: r_{Itent}, r_{React}). We utilize three T5 encoders for encoding input context, the user's causality information, and the system's causality information.

$$
\begin{aligned}
z_c &= T5_{enc}^{c}(c), \\
z_{user} &= T5_{enc}^{user}(k_{user}), \\
z_{sys} &= T5_{enc}^{sys}(k_{sys}).
\end{aligned}
\tag{4.1}
$$

Emotion Classification In order to detect the user's affective state, we concatenate the context representations and the user's causality information, and then pass them through a linear layer followed by a softmax operation to produce the emotion category distribution:

$$p_e = \text{softmax}(W_e([z_c; z_{user}])), \tag{4.2}$$

where W_e is the weight vector of the linear layer. Given the ground-truth emotion label e^* for each conversation, the cross-entropy loss is computed to optimize the process of emotion classification:

$$\mathscr{L}e = -\log(p_e(e^*)). \tag{4.3}$$

Response Generation We cancatenate and feed the information of the user's context and the corresponding causality explanation of the user and the system to a fully-connected (FC) layer.

$$z_{fused} = \text{FC}([z_c; z_{user}; z_{sys}]). \tag{4.4}$$

Subsequently, the target response r_{T5} = [y_1,...,y_T] with length T, is generated by the T5 decoder token by token:

$$p\left(y_t | c_{k_{user},k_{sys}}, y_{<t}\right) = T5_{dec}^{c}(E_{y<t}, z_{fused}), \tag{4.5}$$

where $E_{y<t}$ denotes the embeddings of the tokens that have been generated. The negative log-likelihood for generation is defined as:

$$\mathscr{L}_{gen} = -\sum_{t=1}^{T} \log p\left(y_t | c_{k_{user},k_{sys}}, y_{<t}\right). \tag{4.6}$$

The combined loss is defined as:

$$\mathscr{L} = \mathscr{L}_e + \mathscr{L}_{gen}. \tag{4.7}$$

4.4 Evaluation of Causality Explanation based on ChatGPT

We first evaluated how the output of the causality reasoning module is matched with the reaction and intention of the actual (ground-truth) response.

4.4.1 Setting

Our experiments were conducted on the English EMPATHETICDIALOGUES (ED) dataset. For the experiments based on ChatGPT, we used the "gpt-3.5-turbo" engine version with a temperature of 0. We used 10% of the ED test set for this evaluation (250 samples for single-turn and multi-turn settings, respectively).

4.4.2 Automatic Metrics

(Macro-averaged) F1 score [130], precision, and recall are computed by matching the portion of words in the generation and ground truth that overlap after removing stopwords.
BLEU [121] evaluates the matching between n-grams of the generated response to the ground truth. We utilize BLEU-2, BLEU-3, and BLEU-4 scores.
BERTScore [124] is a BERT-based evaluation measure for text generation, which focuses on lexical semantic similarity between the generated response and the ground truth. We adopt its precision, recall, and F1 score (PBERT, RBERT, FBERT). We used the RoBERTa-Large [50] version.

4.4.3 Case Analysis on the COMET

We evaluated the effectiveness of COMET in inferring intents and reactions since ChatGPT's ability to reason them is sensitive to the given in-context examples. We assessed 60 samples from the ED dataset based on two evaluation metrics: (1) Whether the inferred intents or reactions capture the context; (2) whether there are any conflicts among the generated intents or reactions. We found that 51 out of 60 intent predictions and 46 out of 60 reaction predictions were acceptable. Tables 4.3 and 4.4 show the example of reasoned intentions and reactions, respectively.

Table 4.3 Example intents inferred from COMET.

An accepted example: sys: Did you suffer any injuries? sys's intents: to make sure they are ok; to know if you are ok.
An unaccepted example that does not satisfy metric (1) sys: I understand that one, they are my favorite place to eat. sys's intents: to eat food; to eat good.
An unaccepted example that does not satisfy metric (2) sys: Jeez! It's so unfortunate... very sad really. sys's intents: to be sad; to be happy.

Table 4.4 Example reactions referred by COMET.

An accepted example sys: That's not good. Do you own a gun? sys's reactions: scared; worried; nervous; fearful; angry
An unaccepted example that does not satisfy metric (2) sys: oh man. I'm all about discipline! I don't like spoiled bratty kids. sys's reactions: angry; good; happy; controlling; bad

4.4.4 Results and Analysis

Then we evaluated the performance of the system's intention/reaction reasoning under a different number of in-context examples. Experimental results in Table 4.5 show that increasing the value of *k* allows for ChatGPT to generate reactions and intentions that are more closely aligned with those inferred by COMET from the ground truth response.

Table 4.5 Evaluations of reaction and intention reasoned by ChatGPT+Causality$_{user,sys}$, and we set the corresponding knowledge of ground-truth response inferred by COMET as the reference. PBERT, RBERT, and FBERT represent BERTScore in terms of precision, recall, and F1, respectively.

	Reaction							Intention						
k	F1	BLEU-2	BLEU-3	BLEU-4	PBERT	RBERT	FBERT	F1	BLEU-2	BLEU-3	BLEU-4	PBERT	RBERT	FBERT
2	19.32	6.81	3.16	1.56	91.92	92.60	92.25	13.29	14.65	6.39	3.49	88.90	89.17	89.02
3	21.83	7.12	3.25	1.34	92.28	92.74	92.50	14.49	17.39	8.91	5.37	89.13	89.40	89.26
4	25.83	8.74	3.72	1.48	92.55	92.92	92.73	15.14	19.05	10.07	6.14	89.30	89.54	89.41
5	27.87	8.52	3.55	**1.69**	92.76	92.95	92.85	15.00	19.74	10.69	6.51	89.29	89.46	89.37
6	**29.53**	**9.43**	**4.14**	0.00	**93.15**	**93.22**	**93.18**	**15.71**	**20.72**	**11.55**	**7.25**	**89.62**	**89.76**	**89.68**

4.5 Evaluations on ChatGPT-Based Response Generation

Next, we evaluated the responses generated by ChatGPT.

4.5.1 Evaluation Models

ChatGPT: The prompt given to ChatGPT includes only the chosen in-context raw examples $\mathscr{S}$ from the training set, along with the test sample.
ChatGPT+Causality$_{user,sys}$: The commonsense-based causality explanation prompt $\mathscr{M}^{+}_{prompt}$ is utilized to generate a response by ChatGPT, as illustrated in Algorithm 1.

4.5.2 Evaluation Metrics

Automatic Metrics

EMOACC: Following Welivita and Pu [1] and Lee et al. [90], we utilized the EMOACC[2] to measure the emotion accuracy of the generated responses, which is a BERT-base [48] model finetuned on the English ED dataset.
EMPTOME [85]: It consists of three empathy metrics: **Interpretations (IP)**, which represent expressions of acknowledgments or understanding of the interlocutor's emotion or situation. For example, a response like *"I also worked hard for the math exam, which made me anxious,"* is considered a stronger interpretation than *"I understand how you feel."* **Explorations (EX)**, which represent expressions of active interest in the interlocutor's situation. For instance, a statement like *"Are you feeling terrified right now?"* exhibits stronger exploration compared to *"What happened?"* **Emotional Reactions (ER)**, which represent expressions of explicit emotions. They are computed by pretrained empathy identification models.[3] Specifically, RoBERTa [50] models are separately finetuned for each metric by evaluating the generated response to the number of 0, 1, or 2, a higher value means stronger empathy.
Coherence: We leveraged BERTScore [124] to quantify coherence by computing the semantic similarity between the generated response and the input context.

Human A/B Test

We also conducted an A/B test to compare the performance of *ChatGPT+Causality*$_{user,sys}$ and *ChatGPT*. For each comparison, three crowd-workers were asked to choose the better

[2] https://github.com/passing2961/EmpGPT-3
[3] https://github.com/behavioral-data/Empathy-Mental-Health

Table 4.6 Evaluations on varying the number of in-context examples k in the prompt.

	EMOACC	IP	EX	ER
k=2	0.24	0.08	**0.57**	**1.10**
k=3	0.25	0.09	0.48	1.05
k=4	**0.27**	0.09	0.40	1.04
k=5	0.25	**0.10**	0.33	1.00
k=6	0.25	0.08	0.32	1.01

Table 4.7 Evaluations on the effectiveness of causality$_{user,sys}$ when k set to 2 and 4 with the single-turn setting for our ChatGPT-based methods.

	Method	Empathy				Coherence		
		EMOACC	IP	EX	ER	PBERT	RBERT	FBERT
k=2	ChatGPT	0.060	0.073	0.341	0.923	0.877	0.872	0.875
	ChatGPT+Causality$_{user,sys}$	**0.280**	**0.104**	**0.768**	**1.116**	**0.886**	**0.878**	**0.882**
k=4	ChatGPT	0.036	0.081	0.323	0.867	0.882	**0.875**	0.879
	ChatGPT+Causality$_{user,sys}$	**0.280**	**0.120**	**0.528**	**1.076**	**0.888**	0.874	**0.881**

one or select “Tie” based on three aspects: Empathy, Coherence, and Informativeness [80]. (1) **Empathy (Emp.)** measures whether the generated response understands the user’s feelings and experiences. (2) **Coherence (Coh.)** measures whether the response is coherent/relevant in context. (3) **Informativeness (Inf.)** evaluates whether the generated response conveys more information corresponding to the context.

4.5.3 Results and Analysis

Number of In-context Examples

We investigated the effect of the number of in-context examples using our proposed commonsense-based causality explanation prompt. Table 4.6 shows that setting k to 4 results in the highest emotion accuracy, and setting k to 2 yields better exploration and emotional reactions. Therefore, we selected k values of 2 and 4 for the experiments.

Experimental Results

Tables 4.7 and 4.8 present the results of *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$ with k set to 2 and 4, under the single-turn and multi-turn settings, respectively. In the single-turn setting, a test sample consists of one utterance, while in the multi-turn setting, a test sample contains

Table 4.8 Evaluations on the effectiveness of causality$_{user,sys}$ when k set to 2 and 4 with the multi-turn setting for our ChatGPT-based methods.

	Method	Empathy				Coherence		
		EMOACC	IP	EX	ER	PBERT	RBERT	FBERT
k=2	ChatGPT	0.083	**0.065**	0.318	0.917	0.891	0.902	0.894
	ChatGPT+Causality$_{user,sys}$	**0.199**	0.058	**0.397**	**1.094**	**0.899**	**0.907**	**0.901**
k=4	ChatGPT	0.062	**0.072**	**0.297**	0.866	0.896	0.904	0.898
	ChatGPT+Causality$_{user,sys}$	**0.256**	0.065	0.282	**1.007**	**0.902**	**0.904**	**0.901**

Table 4.9 Human A/B test when k set to 2 and 4 with the single-turn setting for our ChatGPT-based methods.

Comparisons	Aspects	Win	Loss	Tie
ChatGPT+Causality$_{user,sys}$ vs. ChatGPT (k=2)	Emp.	**50.7**	36.0	13.3
	Coh.	**42.7**	42.0	15.3
	Inf.	**51.3**	37.3	11.3
ChatGPT+Causality$_{user,sys}$ vs. ChatGTP (k=4)	Emp.	**49.3**	32.7	18.0
	Coh.	20.0	24.0	**56.0**
	Inf.	**43.3**	40.7	16.0

multiple turns. From the four comparisons, we observed that *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* in at least 5 out of 7 evaluation metrics. Notably, *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* on *EMOACC* and *ER*, indicating that *ChatGPT+Causality*$_{user,sys}$ can generate responses with appropriate emotions. This can be attributed to the inclusion of inferred user emotions and reasoned system emotions, which provide appropriate affective information for generating empathetic responses. This improvement addresses the limitation of *ChatGPT* on emotion recognition, as highlighted in Zhao et al. [21].

ChatGPT+Causality$_{user,sys}$ performs better when k is set to 2 under the single-turn setting. Overall, the performance of *ChatGPT+Causality*$_{user,sys}$ is superior in the single-turn setting compared to the multi-turn setting. This discrepancy can be attributed to COMET, which is trained based on events, not context, making it less effective in predicting causality for long context. Addressing the limitations of COMET will be a focus of our future work.

The results of the human A/B test in Table 4.9 show that *ChatGPT+Causality*$_{user,sys}$ is better than *ChatGPT* on the aspects of *Empathy* and *Informativeness* because of the enriched knowledge by the commonsense-based causality explanations.

Table 4.10 Automatic evaluation results of baselines and our T5-based method. Bold denotes the best score.

	Methods	PPL ↓	BLEU-2	BLEU-3	BLEU-4	D1	D2	PBERT	RBERT	FBERT
Baselines	MOEL	37.63	8.63	4.25	2.43	0.38	1.74	86.19	85.67	85.91
	MIME	36.84	8.37	4.31	2.51	0.28	0.95	86.27	85.59	85.92
	EmpDG	38.08	7.74	4.09	2.49	0.46	1.90	86.09	85.49	85.78
	CEM	36.36	6.35	3.55	2.26	0.54	2.38	86.61	85.39	85.98
	LEMPEx	30.42	2.1	0.8	0.35	1.02	**10.81**	83.60	83.09	83.34
Ours	T5	46.13	3.59	1.94	1.15	0.49	2.82	86.69	84.07	85.35
	T5+Causality$_{user}$	15.26	4.84	1.97	0.89	**1.08**	10.75	90.16	89.48	89.80
	T5+Causality$_{user,sys}$	**13.07**	**10.53**	**6.34**	**4.06**	0.75	5.52	**92.24**	**90.76**	**91.48**

Table 4.11 Results of human A/B test for our T5-based model.

Comparisons	Aspects	Win	Loss	Tie
T5+Causality$_{user,sys}$ vs. CEM	Emp.	**42.0**	40.0	18.0
	Coh.	**38.7**	33.3	28.0
	Inf.	38.3	**44.3**	17.3
T5+Causality$_{user,sys}$ vs. LEMPEx	Emp.	**53.0**	35.0	12.0
	Coh.	**39.0**	33.3	27.7
	Inf.	**50.0**	38.0	12.0

4.6 Experiments on T5-Based Response Generation

Finally, we evaluated the responses generated by the T5-based model.

4.6.1 Evaluation Metrics

(1) Perplexity (PPL) [120], which measures the confidence in the generated response. (2) BLEU (see 4.4.2). (3) D1/D2 (Distinct-1/ Distinct-2) [123], which measure the diversity of responses by calculating the ratio of unique unigrams (D1) and bigrams (D2) to the total number of unigrams and bigrams in the generated responses, respectively. (4) BERTScore (see 4.4.2). (5) Human A/B Test.

4.6.2 Evaluation Models

Affection-based Methods: MoEL [73]: This is an extension of Transformer, which softly combines multiple emotion-specific decoders to a meta decoder to generate an empathetic response. MIME [74]: This method integrates emotion grouping, emotion mimicry, and stochasticity into the emotion mixture to generate diverse empathetic responses. Em-

Table 4.12 Evaluation results of the responses generated by our T5-based method and baselines. The closest to the ground truth is marked as bold.

Methods	EMOACC	IP	EX	ER
MoEL	0.103	0.184	0.209	1.166
MIME	0.076	0.099	0.207	1.256
EmpDG	0.091	0.150	0.169	1.270
CEM	0.091	0.091	0.569	0.950
LEMPEx	0.090	0.135	0.861	**0.575**
T5	0.049	0.110	0.408	1.299
T5+Causality$_{user}$	0.093	0.172	**0.685**	0.784
T5+Causality$_{user,sys}$	**0.125**	**0.271**	0.498	0.751
Ground Truth	0.190	0.279	0.688	0.501

pDG [125]: This model detects nuanced emotions and integrates them into the decoder. And it employs an emotional discriminator and a semantic discriminator to incorporate user feedback.

COMET-based Method: CEM [80], which employs commonsense knowledge, such as the user's reactions, intentions, desires, needs, and effects, to enhance its understanding of the interlocutor's situations and emotions.

T5-based Method: LEMPEx [87], which adopts T5 as the encoder-decoder and utilizes a combination of exemplar-based retrieval, a response generator, and an empathy control module to generate empathetic responses.

T5 [53]: We utilize the T5 model as our base encoder-decoder architecture, integrating with the emotion classifier. We train it from scratch on the EmpatheticDialogue dataset.

T5+Causality$_{user}$: The T5 model is extended with an additional T5 encoder for user's desires/reactions.

T5+Causality$_{user,sys}$: The T5 model is extended with two T5 encoders for the user's causality attributes (desires/reactions) and the system's causality attributes (intentions/reactions), respectively.

4.6.3 Settings

We trained T5-small [53] from scratch on the English ED dataset. We set the learning rate to 0.00001, the batch size to 8, and utilized the top-*k* search decoding strategy with *k* set to 20, sampling with a temperature of 0.2, and a maximum generation length of 40.

Table 4.13 Automatic evaluation results of T5+Causality$_{user,sys}$ and ChatGPT+Causality$_{user,sys}$ (k=2, with whole test set and both single and multi-turn settings).

Evaluations		T5+ Causality$_{user,sys}$	ChatGPT+ Causality$_{user,sys}$
Empathy	EMOACC	0.125	**0.235**
	IP	**0.271**	0.046
	EX	0.498	**0.668**
	ER	0.751	**1.109**
Diversity	D1	0.75	**2.91**
	D2	5.52	**16.44**
BLEU	BLEU-2	**10.53**	3.95
	BLEU-3	**6.34**	2.17
	BLEU-4	**4.06**	1.32

4.6.4 Results and Analysis

Previous studies [80, 87] have shown that CEM and LEMPEx outperformed MoEL, MIME, and EmpDG. Therefore, we compared our method with CEM and LEMPEx in the human A/B test. Automatic evaluation results shown in Table 4.10 and human A/B test results shown in Table 4.11 demonstrate the effectiveness of the proposed commonsense-based causality explanation (Causality$_{user,sys}$). The performance comparison presented in Table 4.12 demonstrates the superiority of our method over the baselines in terms of emotion accuracy (EMOACC), interpretation (IP), and emotion reaction (EX) when compared to the ground truth.

4.6.5 Comparison between T5-based and ChatGPT-based Response Generation

We conducted a performance comparison between the T5-based and ChatGPT-based response generation, as presented in Table 4.13. In terms of "Empathy", *ChatGPT+Causality*$_{user,sys}$ outperforms *T5+Causality*$_{user,sys}$ for EMOACC, EX, and ER, but performs worse for IP. Stronger interpretation (IP), which involves understanding and empathizing through shared experiences [85], is more frequently observed in the T5-based model, which was trained from the ground truth. In contrast, ChatGPT-based generation is not constrained by the ground truth and tends to respond from the perspective of a machine.

In terms of "Diversity" and "BLEU," it is evident that *ChatGPT+Causality*$_{user,sys}$ exhibits a larger diversity but results in a higher degree of mismatch with the ground truth (lower

BLEU scores), indicating a potential need of balancing the response diversity and the accuracy in generating empathetic responses.

4.7 Case Studies and Error Analysis

Table 4.14 shows a case about the comparison between *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$, and illustrates the impact of our proposed commonsense-based causality explanation. We can see that both the responses by *ChatGPT* and *ChatGPT+Causality*$_{user,sys}$ show emotion reactions to the user’s context. However, *ChatGPT+Causality*$_{user,sys}$ outperforms *ChatGPT* by providing detailed suggestions that align with the user’s desires based on reasoned intentions. As discussed in Section , COMET is not always reliable in its predictions. This sensitivity is evident in Table 4.15, where the user’s inferred desires mislead the reasoned intentions of the system.

Table 4.16 further shows comparative case studies between T5-based and ChatGPT-based models with corresponding baselines.

Table 4.14 Case studies of the proposed method and baseline. Blue texts represent cognition, and orange texts show affection. Desires$_{user}$ and Reactions$_{user}$ are inferred by COMET, while Intentions$_{sys}$ and Reactions$_{sys}$ are reasoned by ChatGPT+Causality$_{user,sys}$. “sys” is the abbreviation of system.

Context	user: I’m so scared of bugs! i found one in my hair yesterday and almost died.
ChatGPT	sys: That’s so scary! Did you manage to get it out?
Desires$_{user}$	user wants: to get rid of bugs; to go to the doctor; to run away.
Reactions$_{user}$	user reacts to: scared; frightened; terrified; afraid; nervous.
Intentions$_{sys}$	sys’s intent: to comfort; to help; to offer advice; to be supportive; to avoid bugs.
Reactions$_{sys}$	sys reacts to: empathetic. caring. helpful. understanding. supportive.
ChatGPT+Causality$_{user,sys}$	sys: I know how you feel, bugs are the worst. Have you tried using bug spray or wearing a hat outside?

Table 4.15 Error analysis of COMET knowledge’s negative impact on the causality explanation.

Context	user: I’ m so grateful for my family due to an incident with my friend.
ChatGPT	sys: What happened with your friend?
Desires$_{user}$	user wants: to make amends to their friend; to make amends for the incident.
Intentions$_{sys}$	sys’s intent: to help; to support; to be a good friend; to be there for user.
ChatGPT+Causality$_{user,sys}$	sys: I’m here for you if you need any help making amends with your friend.

4.8 Conclusions and Future Work

We have proposed a commonsense-based causality explanation approach for diverse empathetic response generation that considers the system's intentions and reactions as well as the user's desires and reactions. Specifically, we enhance ChatGPT's ability to reason the system's intentions and reactions by integrating in-context learning with commonsense knowledge (desire, reaction, and intention). We have integrated the commonsense-based causality explanation with both ChatGPT and a trained T5 model. The experimental results demonstrate that our method outperforms other competitive methods on both automatic and human evaluations.

In the future, we will explore fine-grained approaches for causality explanation from the perspective of both the user and the system.

Table 4.16 Comparative case studies between T5-based and ChatGPT-based models with corresponding baselines. Causality$_{user,sys}$ includes both Causality$_{user}$ and Causality$_{sys}$.

	Context	user: I'm so sad because I've read an article about a newborn girl who died because her parents didn't believe in medication and doctors.	
Baselines	MOEL	sys: I am so sorry to hear that. I hope you are able to find her.	
	MIME	sys: That is sad. I am sorry to hear that.	
	EmpDG	sys: I am sorry to hear that. I hope you find out of it.	
	CEM	sys: Oh no, I am so sorry to hear that.	
	LEMPEx	sys: I'm sorry to hear that. I'm so sorry to hear that.	
Ours	T5		sys: Oh no! I'm so sorry to hear that. What happened?
	T5+Causality$_{user}$	Causality$_{user}$	user wants: to tell their friends; to cry; to find out what happened. user reacts to: sad; sorrowful; depressed; upset.
		Response	sys: Oh, that's terrible. I am so sorry for your loss.
	T5+Causality$_{user,sys}$	Causality$_{sys}$	sys's intent: to comfort spk1; to show empathy; to be supportive. sys reacts to: sad; upset; empathetic; sympathetic; concerned.
		Response	sys: I know what you mean, I would have been so mad!
Baseline	ChatGPT	sys: That's heartbreaking. It's important to trust medical professionals and seek help when needed.	
Ours	ChatGPT+Causality$_{user,sys}$	sys: That's really heartbreaking. I'm here for you if you need to talk or just need someone to listen.	
	Context	user: I think that the girl of my dreams likes somebody else. I feel very sad about it.	
Baselines	MOEL	I am sorry to hear that. I hope you can find a new one.	
	MIME	I am sorry to hear that. I hope you can find a good relationship with her.	
	EmpDG	I am sorry to hear that.	
	CEM	I am sorry to hear that.	
	LEMPEx	I understand that. it's really stressful.	
Ours	T5		I'm sorry to hear that. I hope you can find a new job.
	T5+Causality$_{user}$	Causality$_{user}$	user wants: to find a new girlfriend; to ask her out. user reacts to: sad; upset; unhappy; disappointed; angry.
		Response	sys: what did he buy you?
	T5+Causality$_{user,sys}$	Causality$_{sys}$	sys's intent: to comfort; to encourage; to help find a new love. sys reacts to: sympathetic; caring; supportive; encouraging.
		Response	sys: I know how you feel. I have been there before and it is hard to find a new one.
Baseline	ChatGPT	sys: I'm sorry to hear that. It's tough when we have feelings for someone who doesn't feel the same way.	
Ours	ChatGPT+Causality$_{user,sys}$	sys: I'm sorry to hear that. Maybe it's time to move on and find someone who appreciates you.	

Chapter 5

Improving User Personality Recognition in Dialogue

5.1 Introduction

Personality recognition aims to identify an individual's characteristic patterns of feeling, thinking, and behaving, which make each different from one another [131]. Such capability is essential in the realm of human-robot interaction, where correctly detecting the user's personality can significantly enhance the robot's ability to tailor user-adaptive responses, thus fostering richer and more effective human-robot dialogues. Big Five traits [15], which encompass the dimensions of Openness, Neuroticism, Extraversion, Agreeableness, and Conscientiousness, and their respective counter-traits are commonly used in the community for personality assessment. In contrast to prior approaches that infer personality traits from self-reported essays [132–134], first impressions [135, 136], or social media activities [137], this study focuses on the extraction of personality traits from dialogue [138–141, 101].

However, the lack of data is a major obstacle because annotating dialogue-level data with personality information is expensive and time-consuming. Each dialogue involves two participants and personality traits are obtained through psychology questionnaires. Thus, we investigate a data augmentation approach. While previous data augmentation studies focus on generating sentence-level data *invariants* [24–27] without corresponding labels, in this study, we generate both the synthetic dialogue data and corresponding synthetic personality traits through the proposed data interpolation method, which fuses two existing data points controlled by a *continuous ratio variable*.

Additionally, accurately modeling both the inter-dependencies between context and interlocutors, as well as the intra-dependencies within speakers in dialogues, remains a

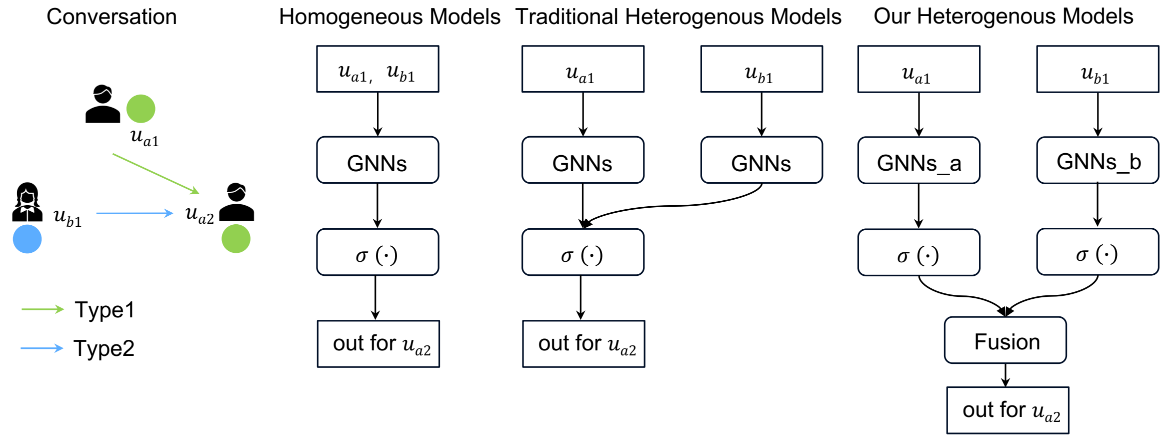


Fig. 5.1 Homogeneous and different heterogeneous models. u_{a1}, u_{a2}, u_{b1} represents alternant utterance of *speakers a* and *b*. $\sigma(\cdot)$ represents activation function.

significant challenge. Previous homogeneous models, such as the graph attention network [28] [29], did not consider the variations in link types. Heterogeneous models like relational graph networks (RGCN) employ distinct relation types to model various dependencies. Yet, they utilize shared coefficients across all relation types, which may fail to capture the unique attributes of each relation type, as shown in Fig 5.1. To address this issue, we propose a method to independently model heterogeneous conversational interactions, capturing both contextual influences and inherent personality traits. Our main contributions are as follows:

- We propose a data augmentation method for personality recognition by interpolation from any two existing data points.
- We propose a heterogeneous conversational graph network (HC-GNN) to independently model both the interdependencies among interlocutors, as well as the intra-dependencies within the speaker in dialogues.
- Experimental results using the RealPersonaChat dataset demonstrate that increasing speaker diversity significantly improves personality recognition in both monologue and dialogue settings. The proposed HC-GNN method outperforms baseline models, showcasing its effectiveness.

5.2 Related Work

This section introduces related work in three key areas: personality recognition in dialogues, data augmentation, and various graph neural networks.

5.2.1 Personality Recognition in Dialogue

Mehl et al. [141] pioneered the automatic personality assessment of all Big Five personality traits using various psycholinguistic attributes. To this end, they analyzed a collection of daily-life conversations by 96 participants over 2 days. However, it only contains the subjects' conversation; we also want to analyze how interlocutors impact the subject's personality expression in the dialogue. Jiang et al. [139] collected a dialogue-based personality dataset, `FriendsPersona`, by annotating five personality traits of speakers from Friends TV Show through crowdsourcing. Chen et al. [140] collected a large-scale Chinese personalized and emotional dialogue dataset CPED. Nonetheless, the Big Five personality labels were assigned by external observers rather than derived from self-assessments by the speakers. Han et al. [142] created a multi-party conversation-based personality dataset derived from CPED, consisting of 1195 data samples for personality recognition, and introduced a speaker-aware layering named SH-Transformer converter. Most recently, Yamashita et al. [101] presented the RealPersonaChat (RPC) dataset by documenting the authentic personality traits of the participants and allowing them to freely engage in dialogues. This dataset aligns closely with our research objectives, as it provides a foundation for evaluating the personality traits of subjects who may engage in chit-chatting dialogue with a conversational agent. However, this dataset has a relatively limited number of speakers (233). This sparsity poses a challenge in effectively detecting the personality traits of unseen speakers. We propose a data augmentation method to enrich the speaker diversity.

5.2.2 Data Augmentation

Data augmentation (DA) tries to fill the gap between the data distribution of the training set and the real data with no annotation cost. Previous DA studies focus on generating data *invariants*. In the computer vision field, simple geometric transformations like cropping, rotation, and noise injection can be easily applied to continuous image data [143]. Due to the discrete nature of language data, previous DA studies in NLP usually involve discrete noises including *1)* character-level modification like changing character case [24], *2)* subword-level regularization such as BPE-dropout [144], *3)* word-level replacement, insertion, or deletion [25], and *4)* sentence-level modification such as paraphrasing [145, 26] and back-translation [27, 146–148]. Our method differs from them in two aspects. First, we generate data *variants* including both dialogue data and corresponding personality labels. Second, we generate synthetic personality traits following a continuous distribution from existing discrete trait data by introducing a random fusion variable. This bridges the gap between the discrete distribution in the dataset and the continuous distribution in reality. A similar work is the example extrapolation method [149] which generates augmented embeddings of the target domain by leveraging the similarity of embedding spaces from another assisting domain. Different from it, our interpolation method requires only one dataset.

5.2.3 Graph Neural Networks

Compared to Convolutional Neural Networks, which are designed to process grid-like data [150–154], Graph Neural Networks (GNNs) can handle data structured as graphs. Graph Neural Networks (GNNs) and their variants have been widely applied in dialogue-related tasks, like conversational emotion recognition [155–157], and dialog act classification [158]. This is primarily due to the adjacency matrix in GNNs effectively simulating interactions within conversations. Yang et al. [159] proposed a dynamic deep graph convolutional network for personality detection on social media posts. Existing methodologies, whether employing static or dynamic approaches to construct interactions within graphs, mainly focus on homogeneous or heterogeneous conversation modeling. Nevertheless, various types of nodes and links have different traits and their features may fall in different spaces. For instance, as illustrated in Fig 5.1, traditional heterogeneous models like RGCN [160] utilize shared coefficients across all relation types, potentially failing to capture the unique attributes of each relation type. To solve it, this paper proposes a modification to the existing heterogeneous model framework. We introduce separate GNNs to distinctly capture the diverse relation types, thereby respecting the unique properties of each node and link type.

5.3 Proposed Method

5.3.1 Data Interpolation

This section describes math notations, how to fuse two existing data points to generate synthetic dialogue and Big Five traits, and variants of the proposed method.

Notations. Each dialogue D contains utterances u_a from the *speaker a* or u_b from the *speaker b* in alternant turns, that is $D = \{u_{a1}, u_{b1}, u_{a2}, u_{b2}, ..., u_{an}, u_{bn}\}$, where n is the number of turns. Each dialogue is accompanied by a label $\mathbf{y}$ that is a vector containing the Big Five personality traits of the target speaker. We aim to generate synthetic dialogue D_{syn} and its label $\mathbf{y}_{syn}$ from two existing dialogues D_1 and D_2 and their labels $\mathbf{y}_1$ and $\mathbf{y}_2$.[1]

Dialogue Interpolation. First, we randomly select two dialogues D_1 and D_2 in the training set. Second, we split each dialogue into chunks (c) each containing t turns, which is a

[1] Since we focus on the personality of the initiating *speaker a* in the experiments, $\mathbf{y}_1$ or $\mathbf{y}_2$ refers to the personality of *speaker a*.

hyper-parameter controlling the context length we desire (we set t=3). This results in $D_1 = \{c_1, c_2, ..., c_l\}$ and $D_2 = \{c'_1, c'_2, ..., c'_l\}$, where $l = \frac{n}{t}$ is the number of chunk one dialogue contains. Finally, we combine chunks from D_1 and D_2 to generate D_{syn} using a fusion ratio β that is a random variable independently sampled from Uniform$(0,1)$ for each synthetic data point. D_{syn} can be represented as:

$$\begin{aligned} D_{\text{syn}} &= \{c_i^{\text{syn}} \mid 1 \leq i \leq l\}, \text{where} \\ c_i^{\text{syn}} &= \begin{cases} c_i & \text{with probability } \beta, \\ c'_i & \text{with probability } 1-\beta. \end{cases} \end{aligned} \tag{5.1}$$

Specially, when generating synthetic monologue data, we split each monologue D into utterances instead of chunks.

Label Interpolation. Because each label is a vector of real numbers represented as $\mathbf{y} \in \mathbb{R}^5$, we can simply obtain the synthetic label through:

$$\mathbf{y}_{syn} = \beta \mathbf{y}_1 + (1-\beta)\mathbf{y}_2. \tag{5.2}$$

Method Variants. There are three types of variants of the proposed method (former setting used). First, we can sample $\beta \sim$ Uniform$(0,1)$ or fixing β to 0.5. The former setting produces a richer variety of data. Second, we can either select two dialogues possibly from different speakers, where $\mathbf{y}_1$ and $\mathbf{y}_2$ independently and identically distributed ($\mathbf{y}_1 \overset{\text{iid}}{\sim} \mathbf{y}_2$), or select two dialogues from the same speaker $\mathbf{y}_1 = \mathbf{y}_2$ which results in $\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y}_{syn}$. The former setting can produce new speakers with synthetic personality traits. Third, the length of the synthetic dialogue can be sampled from Uniform$(t_{min}, |D_{syn}|)$ by truncating (we set $t_{min} = 2$), or equal to $|D_{syn}|$. This enables personality recognition in early turns which is preferred in real applications.

5.3.2 Heterogeneous Conversational Graph Neural Network

We map each utterance in the dialogue into embeddings. Subsequently, we describe heterogeneous conversation modeling, followed by an explanation of heterogeneous conversational graph network feature encoding and fusion.

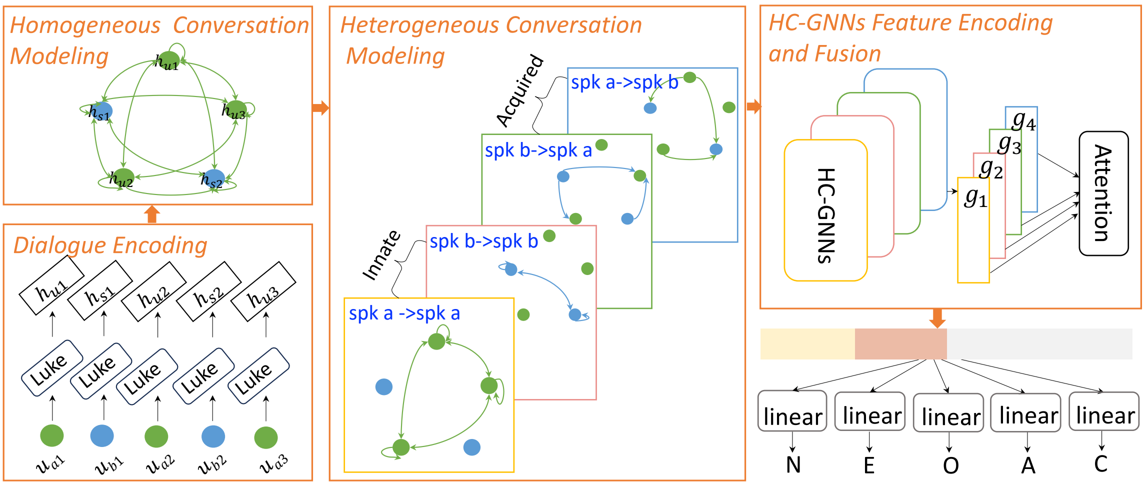


Fig. 5.2 Proposed heterogeneous conversational graph neural network (HC-GNN), which captures the interdependencies among interlocutors (acquired) and the intra-dependencies within *speaker a* or *b* (innate).

Dialogue Encoding

For each dialogue $D = \{u_{a1}, u_{b1}, u_{a2}, u_{b2}, u_{a3}...\}$, we employ a BERT-like model Japanese Language Understanding with Knowledge-based Embeddings (LUKE) [161] [2] to encode each utterances in the dialogue:

$$\mathbf{h}_{u_i} = LUKE(u_i) \in \mathbb{R}^{1\times d}, \tag{5.3}$$

where $\mathbf{h}_{\mathbf{u_i}}$ denotes the final hidden state of the “[CLS]” token to represent the meaning of the whole utterance, and d is the dimension of the output.

Heterogeneous Conversation Modeling

To explicitly model the interaction between speakers, we independently model the intra-dependency (innate personality) and interdependency (acquired personality which is influenced by the interlocutor), as shown in Fig. 5.2. We denote directed and labeled multi-graphs as $\mathscr{G}_n = (\mathscr{V}_n, \mathscr{E}_n, r)$ with nodes $\vartheta_{n,i} \in \mathscr{V}_n$ and labeled edges (relations) $(\vartheta_{n,i}, r, \vartheta_{n,j}) \in \mathscr{E}_n$, where $r \in \mathscr{R}$ represents one of the conversation relation types $\{spka \rightarrow spka$, $spkb \rightarrow spkb$, $spkb \rightarrow spka$, $spka \rightarrow spkb\}$, n represents the number of graphs.

[2] https://huggingface.co/studio-ousia/luke-japanese-base

HC-GNNs Feature Encoding and Fusion

For each relation type in each graph $\mathscr{G}$, we then encode the features with dynamic attention and graph attention networks [29] to aggregate the interactions between each group of speakers (self and interlocutor):

$$\mathbf{h}_i^{n(l)} = \sigma(\sum_{k\in K}\sum_{j\in N_i^r}\frac{a_{i,j}^{(k)}}{N_i^r}\mathbf{W}_r^{(l)}\mathbf{h}_{u_j} + a_{i,i}^{(k)}\mathbf{W}_0^{(l)}\mathbf{h}_{u_i}), \tag{5.4}$$

where N_i^r denotes the neighboring indices of node i under relation $r \in \mathscr{R}$, K represents the number of attention head. $\mathbf{W}_r^{(l)}$ and $\mathbf{W}_0^{(l)}$ are the learnable weight metrics, l is the layer of HC-GNN, and $\sigma(.)$ is an ReLU activation function. The attention scores are normalized across all neighbors $j \in N_i^r$ using softmax, and the attention function is defined as:

$$\begin{aligned} a_{i,j} &= \text{softmax}_j(e(\mathbf{h}_{u_i},\mathbf{h}_{u_j})) = \frac{\exp(e(\mathbf{h}_{u_i},\mathbf{h}_{u_j}))}{\sum_{j'\in N_i^r}\exp(e(\mathbf{h}_{u_i},\mathbf{h}_{u_{j'}}))}, \\ e(\mathbf{h}_{u_i},\mathbf{h}_{u_j}) &= \mathbf{a}^\top \text{LeakyReLU}(\mathbf{W}\cdot[\mathbf{h}_{u_i};\mathbf{h}_{u_j}]). \end{aligned} \tag{5.5}$$

Here $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{W} \in \mathbb{R}^{d' \times d}$ are learned, ; denotes vector concatenation. We use a graph neural network (GCN) [28] to capture the deeper interaction representations:

$$\mathbf{h}_i^{n(l+1)} = \sigma(\sum_{j\in N_i^r}\mathbf{W}^{(l+1)}\mathbf{h}_j^{n(l)} + \mathbf{W}_0^{(l+1)}\mathbf{h}_i^{n(l)}), \tag{5.6}$$

where $\mathbf{W}^{(l+1)}$ and $\mathbf{W}_0^{(l+1)}$ are learnable metrics. Given the latent representation $\mathbf{g}_n$ of each graph which corresponds to a distinct relation, we then use the self-attention mechanism to fuse the graph outputs of innate and acquired relations:

$$\begin{aligned} \mathbf{z} &= [\mathbf{g}_0;\mathbf{g}_1;...;\mathbf{g}_n], \\ \mathbf{z}' &= \text{Attn}(\mathbf{z},\mathbf{z},\mathbf{z}). \end{aligned} \tag{5.7}$$

5.3.3 Personality Recognition with Multi-task Learning

We first analyze the correlations between pairs of personality traits in the dataset, as shown in Table 5.1. The p-value of the Pearson correlation for each pair is less than 0.05, indicating statistically significant relationships. We recognize Openness, Neuroticism, Extraversion, Agreeableness, and Conscientiousness in the manner of multi-task learning using five linear

Table 5.1 Pearson correlation between pairs of Big Five personality traits in the dataset. N, E, O, A, and C represent Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, respectively.

Big-Five Pair	(N, E)	(N, O)	(N, A)	(N, C)	(E, O)	(E, A)	(E, C)	(O, A)	(O, C)	(A, C)
Pearson Correlation	-0.49	-0.23	-0.27	-0.15	0.47	0.46	0.26	0.40	0.15	0.36

(for all pairs $p < .05$)

layers. The i_{th} layer can be represented as:

$$\mathscr{P}_i = \sigma(\mathbf{W}_p^i z' + b_p^i), \tag{5.8}$$

where σ denotes the activation function ReLU, $\mathbf{W}_p^i$ and b_p^i are the learnable weight matrix and bias. We treat them as regression tasks and use the mean absolute error (MAE) as the loss function for model optimization. The loss item of each data sample j is denoted as $l(\mathbf{P}^j, \mathbf{y}^j)$ that is calculated by averaging the loss of five tasks, where $\mathbf{P}^j$ is the vector of 5 predictions and $\mathbf{y}^j$ is the vector of 5 ground truth personality traits. The loss item of one batch containing N data samples is denoted as L. They are calculated as follows:

$$\begin{aligned} l(\mathbf{P}^j, \mathbf{y}^j) &= \frac{1}{5}\sum_{i=1}^{5}\left|\mathscr{P}_i^j - y_i^j\right|, \\ L &= \frac{1}{N}\sum_{j=1}^{N} l(\mathbf{P}^j, \mathbf{y}^j). \end{aligned} \tag{5.9}$$

5.4 Experimental Settings

5.4.1 dataset

We conducted experiments using the RealPersonaChat dataset [101]. We conducted two experimental settings, one based on **monologues** and the other on **dialogues**. The monologue setting focuses on a speaker's own utterances, while the dialogue setting integrates utterances from both the speaker and the interlocutor for personality recognition. In the monologue experiments, we implemented strict speaker splitting to ensure no overlap among speakers across the training, validation, and test sets. This approach meant the model was evaluated on unseen speakers.

In the dialogue experiments, ensuring non-overlap of both speakers in all datasets proved challenging. Therefore, we ensured that only the initiating speaker was non-repeating across datasets, and the model was tasked with predicting only the initiating speaker's personality.

Therefore, we only used relation *spk a→spk a* and *spk b→spk b* in the HC-GNN model. We created training, validation, and test sets by randomly dividing the speakers in an 8:1:1 ratio for 100 times and selecting the split that most closely matched 8:1:1 distribution for monologues and dialogues. We then fixed the split across all experiments.

5.4.2 Models

All the models are based on a BERT-like model which converts utterances into embeddings. In all experiments, we used pooled output from LUKE base model [161] because it showed the best performance in terms of average balanced accuracy among LUKE-base/large, RoBERTa-base/large, xlm-roberta base/large, mdeberta-v3-base models. We experimented with different models after the base model:
MLP: Five linear module joints with regression heads are used to predict each personality. Each regression head contains two linear layers: the first layer maps the embedding from LUKE to a 16-dimensional embedding, and the second layer maps the 16-dimensional embedding to a single output.
Homogeneous Methods: Graph Convolutional Networks (**GCNs**) [162] represent one of the most prevalent methods for handling graph-structured data, particularly in node classification and link prediction tasks; **GATv2** [29] introduces a significant enhancement by transitioning from a static to a dynamic attention mechanism.
Heterogenous Mthods: Relational Graph Convolutional Networks **RGCN** [160], are developed to handle the multi-relational data with heterogenous architecture, as shown in Fig 5.1; Heterogeneous conversation graph neural network **(HCGNN)**, the proposed method, which independently model the interdependencies between context and interlocutors and the intra-dependencies within the speaker.

5.4.3 Training

We used the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and a learning rate of 1×10^{-5}. We used a linear scheduler with warmup step = 150. We used Mean Absolute Error (MAE) criterion because it outperformed the Mean Squared Error (MSE) greatly in terms of balanced accuracy. We set the batch size to 128 for linear models and 32 for graph neural network models, which reach the memory limitation of eight 32G GPUs. We calculated the loss on the validation set after each epoch and applied early stopping when no improvement was observed for 3 epochs.

5.4.4 Evaluation Metrics

We report binary classification accuracy and balanced accuracy together with Pearson correlation and Spearman correlation for regression tasks. The threshold for the binary classification task of each personality trait is set to the median score in the training set. Here are the details of each metric:

Accuracy: a metric that summarizes the performance of a classification task, which is the number of correctly predicted data points out of all the data points.

Balanced Accuracy: arithmetic mean of sensitivity and specificity to deal with imbalanced data.

Pearson Correlation: a correlation coefficient that measures the linear correlation between the predicted personality values and the ground truth.

Spearman Correlation: a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables).

Table 5.2 Accuracy results in monologue setting with original data and augmented dataset. The best result in each column is in **bold**.

Data	Accuracy						Balanced Accuracy					
	N	E	O	A	C	Avg.	N	E	O	A	C	Avg.
Original	66.2	52.6	57.6	52.7	57.9	57.4	59.7	54.3	58.8	50.3	55.0	55.6
+10k	74.3	55.5	56.6	49.9	51.2	57.5	60.4	52.5	55.4	47.8	52.8	53.8
+20k	**74.5**	54.3	55.5	**59.0**	60.1	60.7	**65.9**	53.8	56.0	58.5	48.3	56.5
+50k	60.5	53.0	60.9	58.7	59.8	58.6	60.2	55.0	61.2	**62.0**	52.4	58.2
+500k	62.7	**58.2**	**62.0**	57.9	**65.4**	**61.2**	64.6	**56.0**	**61.3**	60.3	**59.7**	**60.4**

Table 5.3 Correlation results in monologue setting with different data size.

Data	Pearson Correlation					Spearman Correlation				
	N	E	O	A	C	N	E	O	A	C
Original	*.279*	-.040	*.473*	*.105*	*.166*	*.292*	-.014	*.282*	*.104*	*.200*
+10k	*.389*	-.092	*.510*	*.100*	***.200***	*.283*	-.064	*.277*	*.091*	*.180*
+20k	***.492***	.054	***.510***	*.268*	*.113*	***.495***	.063	*.276*	*.261*	*.126*
+50k	*.224*	.067	*.486*	***.387***	*.099*	*.264*	.085	*.378*	***.375***	*.119*
+500k	*.267*	.025	*.459*	*.232*	*.164*	*.333*	.045	***.388***	*.230*	***.203***

(*Italic* means $p < .05$)

5.5 Results and Analysis

5.5.1 Monologue

Main Results. Table 5.2 presents a comparative analysis of the accuracy and balanced accuracy of the data augmentation method at various data sizes. With the addition of augmented data, both accuracy and balanced accuracy show significant improvements, increasing from 57.4% to 61.2% and from 55.6% to 60.4%, respectively. Table 5.3 shows the Pearson and Spearman correlation results. We observed that data augmentation generally improves correlation in most traits, and the impact of augmentation is more pronounced in N and A than in others. We failed to predict personality trait E in any setting. We believe this may be due to our dataset being based on first-meeting spontaneous situations, where people tend not to exhibit extrovert traits explicitly. We also found that different from accuracy and balanced accuracy results, the highest Pearson and Spearman correlations do not always occur at the same data augmentation point.

Results of Data Augmentation Variants.

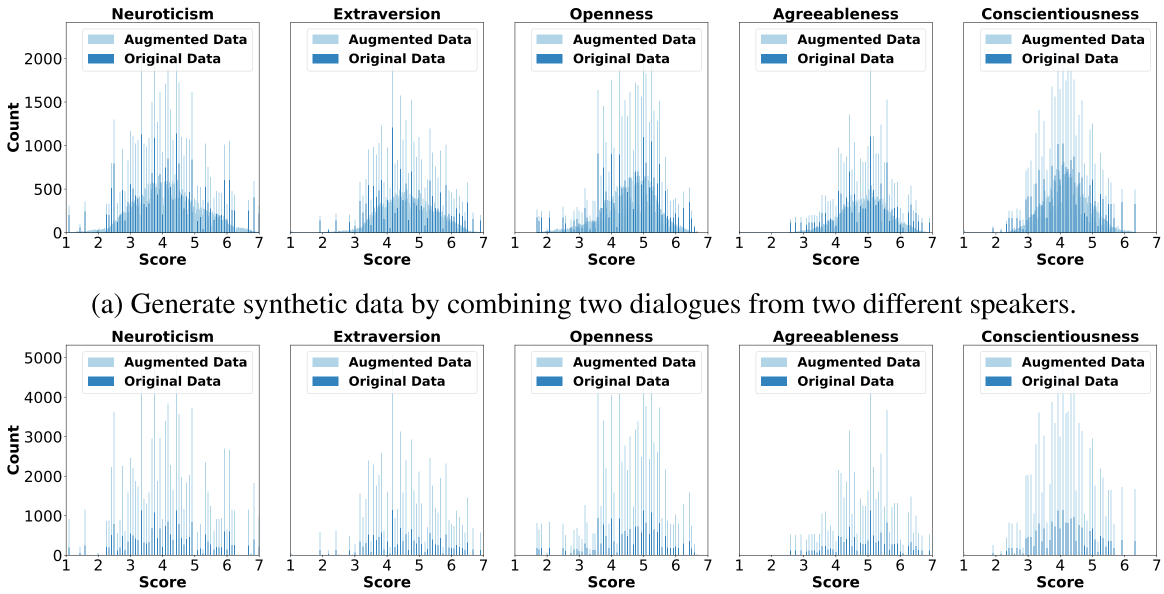
1. Fusing Ratio. We compared results using $\beta \sim \text{Uniform}(0,1)$ and fixed $\beta = 0.5$. With 500k additional data, using random β achieved 61.2% averaged accuracy and 60.4% averaged balanced accuracy whereas using fixed β showed 59.3% accuracy and 58.4% balanced accuracy.

2. Speaker Choice. We compared generating synthetic dialogue from the same speaker or two different speakers. We observed using dialogues from different speakers not only enables continuous data distribution as shown in Fig 5.3 but also showed much higher averaged accuracy (61.2% vs 57.3%) and balanced accuracy (60.4% vs 58.1%), which demonstrates that speaker variety is more crucial than the number of conversations for personality recognition within this dataset.

3. Various Dialogue Lengths. Real-time personality recognition in dialogue is essential for human-robot interaction. We tested this ability by using the first 2 turns of utterances during inference. We found that using various lengths in the augmented data enables 58.2% averaged balanced accuracy that is comparable to the result using the full dialogue 60.4%. However, the result is only 55.5% if we keep all synthetic dialogue full-length (approximately 15 turns).

5.5.2 Dialogue

Comparisons between Monologue and Dialogue. To explore the impact of context on personality recognition, we first appended the [SPK1] or [SPK2] token to the respective



(a) Generate synthetic data by combining two dialogues from two different speakers.

(b) Generate synthetic data by combining two dialogues from the same speaker.

Fig. 5.3 Data Distribution of augmented data and original data.

Table 5.4 Accuracy results in the comparisons among monologue and dialogue.

Model		Accuracy						Balanced Accuracy					
		N	E	O	A	C	Avg.	N	E	O	A	C	Avg.
Monologue	MLP	66.2	52.6	57.6	52.7	57.9	57.4	59.7	54.3	**58.8**	50.3	**55.0**	55.6
Dialogue	MLP	66.8	53.2	**57.9**	49.0	58.4	57.1	59.5	**56.9**	57.9	49.8	47.5	54.3
	GCN [162]	**69.5**	40.3	52.9	33.3	**72.1**	53.6	50.0	50.4	53.7	34.4	50.1	47.7
	GAT [29]	66.4	54.5	52.7	59.9	54.6	57.6	60.4	51.3	50.1	52.2	52.1	53.2
	RGCN [160]	68.5	**57.2**	55.4	41.9	56.7	55.9	**63.0**	54.5	55.9	47.2	51.9	54.5
	HC-GNN (ours)	69.0	53.5	55.6	**65.9**	52.2	**59.2**	60.9	54.3	52.6	**59.6**	54.6	**56.4**

utterances and then concatenated all utterances using the [SEP] token. We used the same model as in the monologue experiment. As indicated in Table 5.4 and 5.5, the results with conventional methods using the context (dialogue) show a decrease in performance compared to the monologue setting across most evaluation metrics. We hypothesize that merely concatenating utterances between two speakers is not an effective method for modeling the interactions between interlocutors. Therefore, we propose independently modeling both the interdependency among speakers and the intra-dependency within the speaker. The results, as shown in Tables 5.4 and 5.5, indicate that our proposed method surpasses all baseline methods in the dialogue setting and marginally improves upon the results in the monologue setting.

Table 5.5 Correlation results in the comparisons among monologue and dialogue.

	Model	Pearson Correlation					Spearman Correlation				
		N	E	O	A	C	N	E	O	A	C
Monologue	MLP	*.279*	-.040	*.473*	*.105*	*.166*	*.292*	-.014	*.282*	*.104*	***.200***
Dialogue	MLP	*.288*	.048	***.492***	*.174*	*.077*	*.214*	.088	*.262*	*.123*	*.072*
	GCN [162]	*.170*	-.015	*.298*	*-.203*	*.079*	*.173*	.017	*.164*	*-.227*	*.105*
	GAT [29]	*.307*	-.067	*.420*	*.079*	*.134*	*.284*	-.030	*.226*	*.082*	*.135*
	RGCN [160]	***.377***	.048	*.490*	*.152*	*.148*	***.375***	.078	***.311***	*.146*	*.160*
	HC-GNN (ours)	*.285*	.040	*.347*	***.216***	***.169***	*.304*	.066	*.243*	***.191***	*.193*

(*Italic* means $p < .05$)

Table 5.6 Accuracy results of HC-GNN in dialogue setting with various data size.

Data	Accuracy						Balanced Accuracy					
	N	E	O	A	C	Avg.	N	E	O	A	C	Avg.
Original	69.0	53.5	55.6	65.9	52.2	59.2	60.9	54.3	52.6	59.6	54.6	56.4
+10k	73.9	59.7	54.0	63.9	54.6	61.2	59.5	52.8	50.7	55.6	59.0	55.5
+20k	**74.3**	**60.1**	58.4	**67.5**	40.8	60.2	60.8	54.4	55.8	61.6	55.5	57.6
+50k	69.6	54.7	58.7	62.2	**63.8**	**61.8**	61.9	**55.5**	**58.4**	**62.3**	55.1	**58.6**
+500k	66.3	56.6	**58.8**	60.1	59.0	60.2	**63.7**	53.8	57.0	57.7	**59.2**	58.3

Data Augmentation in Dialogue. We tested the effectiveness of data augmentation in the dialogue setting. The results, as shown in Tables 5.6 and 5.7, indicate that increasing speaker variety can enhance personality recognition in dialogue. Although the highest balanced accuracy achieved in the dialogue setting is 58.6, falling short of the monologue setting's best result of 60.4. Due to our focus on predicting only *speaker a*'s personality in the dialogue setting, the original dataset lost half of its conversational data for augmentation purposes. This loss is an inevitable trade-off in the pursuit of speaker-independent personality recognition within dialogue settings.

5.6 Conclusion

We have proposed a data augmentation method for personality recognition, which involves interpolating between two existing data points to enhance speaker diversity. Additionally, we have introduced the HC-GNN method to independently model the interdependencies among interlocutors, as well as the intra-dependencies within the speaker in dialogues. Experimental results from the RealPersonaChat dataset demonstrate that increasing speaker diversity

Table 5.7 Correlation results of HC-GNN in dialogue setting.

Data	Pearson Correlation					Spearman Correlation				
	N	E	O	A	C	N	E	O	A	C
Original	*.285*	.040	*.347*	*.216*	*.169*	*.304*	*.066*	*.243*	*.191*	*.193*
+10k	*.369*	.019	*.329*	*.235*	***.234***	*.343*	.058	*.226*	*.234*	***.263***
+20k	*.327*	-.052	***.441***	***.314***	*.110*	*.301*	.003	*.245*	*.278*	*.142*
+50k	*.345*	.030	*.312*	*.300*	*.162*	*.374*	.050	*.250*	***.281***	*.183*
+500k	***.426***	.046	*.411*	*.209*	*.223*	***.439***	.062	***.315***	*.194*	*.242*

(*Italic* means $p < .05$)

significantly improves personality recognition in both monologue and dialogue settings. Our HC-GNN method outperforms baseline models, showcasing its effectiveness. However, our experiments suggest that context did not make a large improvement in personality recognition. Further exploration of the dialogue setting will be the focus of our future work.

Chapter 6

Endowing System with Consistent Personality for Empathetic Response Generation

6.1 Introduction

Empathy and personality are pivotal factors in the development of human-like systems. Empathy is the ability of humans to put themselves in another's position, which encompasses understanding other's experiences and feelings for responding appropriately. Personality is the enduring patterns of thoughts, feelings, and behaviors that distinguish individuals from one another [163].

Empathy integrates cognition and emotion, involving understanding and responding emotionally to others' situations [12]. Consequently, prior research has focused on methods to generate empathetic responses by improving affective expression [73, 74, 125], or exploring context understanding [77, 80, 91, 87]. However, as illustrated in Fig. 6.1, individuals with different personalities can exhibit diverse empathy styles within identical contexts. Previous methods for empathetic response generation did not consider the system's personalities, which leads to responses that may reflect empathy but lack personalization.

Systems that express a consistent personality are important for enhancing believability [30]. As shown in Fig. 6.1, when the system changes its personality in a single conversation, it would make the interaction feel less human-like. Moreover, an appropriate empathetic response may depend on the personality traits. Richendoller and Weaver III [2] examined the relationships between psychoticism, extraversion, and neuroticism and three styles of empathic intents: empathetic, perspective-taking, and sympathetic. Their findings indicate

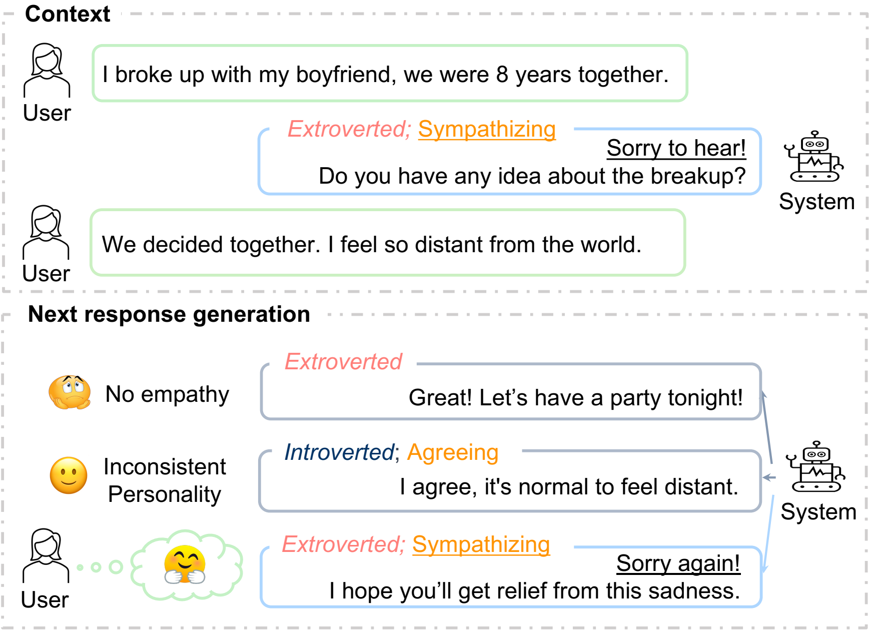

Context
User
I broke up with my boyfriend, we were 8 years together.
Extroverted; Sympathizing
Sorry to hear!
Do you have any idea about the breakup?
System
User
We decided together. I feel so distant from the world.
Next response generation
No empathy
Extroverted
Great! Let's have a party tonight!
Inconsistent Personality
Introverted; Agreeing
I agree, it's normal to feel distant.
System
User
Extroverted; Sympathizing
Sorry again!
I hope you'll get relief from this sadness.

Fig. 6.1 Different *personalities* exhibit distinct preferences for empathetic intents[1] in responses [2, 3]. In a given context, the user shows varying feelings to the system's responses, where the system encompasses empathetic expression and consistent personality traits, resulting in a more human-like interaction.

that individuals with different personalities exhibit distinct preferences for empathetic intents, inspiring our motivation to consider system's personality traits into empathetic response generation. However, the relationship between commonly-used Big Five [15] / Myers-Briggs Type Indicator (MBTI) [16] personalities and empathetic intents has not been fully explored.

To address this problem, we implicitly learn these connections through the prediction of both personality traits and empathetic signals in responses. Empathetic signals include empathetic intents and empathetic communication mechanisms (ECM) [31]. Specifically, ECM includs interpretations (IP), explorations (EX), and emotional reactions (ER). Further inspired by the prefix tuning method employed by Li and Liang [32] and Liu et al. [33], we propose a multi-grained prefix encoder aimed at discerning personality traits alongside empathetic signals.

Because the EMPATHETICDIALOGUES dataset (ED) [60] primarily targets expressing empathy rather than personality, it is hard to learn personality traits from a single response during traditional backpropagation. To address this, we propose a personality enhancement (PE) module that utilizes contrastive learning to calibrate the generation of empathetic responses by integrating explicit personality traits, thereby improving empathy and personality expression in the generated responses. Our main contributions are:

- To the best of our knowledge, this is the first work to consider the system's personality for empathetic response generation. Moreover, we propose a multi-grained prefix mechanism to implicitly learn the relationship between the system's personality and corresponding empathetic expressions.
- We introduce a personality enhancement module to calibrate an empathetic response generation model via contrastive learning for generating responses that are both empathetic and reflective of a distinct personality.

6.2 Preliminaries

Due to the lack of personality and empathetic signal annotations within the benchmark ED dataset, we train distinct models specialized for each aspect.

6.2.1 Personality Predictor

PANDORA [137][2] is the largest dataset of Reddit comments labeled with Big Five and MBTI traits intensities. We strictly partitioned the PANDORA dataset by the user, guaranteeing no user overlap across the training, validation, and test sets. This approach allows us to assess the model's efficacy in identifying the personality traits of unseen users, thereby making the evaluation results on the PANDORA dataset applicable to the ED dataset as well. The Big Five personality trait scores are continuous, ranging from -100 to 100, while MBTI scores are binary. We normalized each Big Five personality trait score to a range between -1 and 1 and balanced the binary labels of each MBTI trait, The details of the statistics are shown in Table 6.1 for reference.

To make the length distribution of the examples similar to the ED dataset, we conducted the following steps for both Big Five and MBTI experiments: 1) only preserved sentences containing ASCII characters with 10 to 50 tokens. 2) For each user we derived non-overlapping samples by randomly selecting and concatenating k sentences, where k was randomly selected to vary between 1 and 5. We incorporated five fully connected layers with ReLU activation followed by five regression heads on top of the LUKE model, to predict all Big Five trait intensities simultaneously. We separately finetuned the LUKE model with one fully connected layer and one regression head for each MBTI trait prediction. For all the experiments, the learning rate was set as 1e-5, the dropout is 0.1, and the mean squared error loss. We used a linear scheduler with a warmup step of 100. Using the median of the

[2] https://psy.takelab.fer.hr/datasets/all/pandora

training label and 0.5 as the threshold, we further binarize the predicted intensities and actual labels and report the accuracies and F1 scores for Big Five and MBTI, separately. Based on the prediction accuracy shown in Table 6.2, we adopt the combination of MBTI introverted, MBTI thinking, and Big Five extraversion as personality traits used in this study.

	Traits	unique	train	valid	test
MBTI	Introverted	speakers	1,531 | 1,402	197 | 170	193 | 174
		utterances	412,467 | 424,008	55,870 | 48,218	49,167 | 56,177
	Intuitive	speakers	820 | 995	100 | 126	106 | 120
		utterances	268,470 | 277,440	38,443 | 30,230	34,022 | 34,527
	Thinking	speakers	2,568 | 1,728	307 | 230	334 | 205
		utterances	547,753 | 561,814	70,483 | 66,916	72,527 | 66,181
	Perceiving	speakers	2,965 | 3,110	388 | 371	392 | 367
		utterances	871,439 | 877,865	109,267 | 108,546	107,740 | 112,082
Big5	All	speakers	1,225	153	154
		utterances	102,523	12,803	12,803

Table 6.1 Statistics of unique speakers and utterances across each MBTI and all Big Five traits in the filtered PANDORA dataset. For MBTI traits, we show the number of label 0 | 1.

	Traits	Acc.	BA.	F1	Pear.	Spear.
MBTI	**Introverted**	**59.11**	**58.15**	**65.41**	***0.1838***	***0.1852***
	Intuitive	50.50	50.39	56.83	*-0.0592*	*-0.0506*
	Thinking	**59.30**	**59.06**	55.79	***0.2344***	***0.2287***
	Perceiving	49.16	49.26	47.00	*-0.0166*	*-0.0157*
Big5	Agreeable	47.72	47.45	0.5468	*-0.0274*	*-0.0312*
	Conscientious	52.46	53.75	0.5663	*0.1291*	*0.1016*
	Extraversion	**67.23**	**63.70**	**0.7566**	***0.4081***	***0.3862***
	Neuroticism	53.91	54.02	0.5696	*0.1074*	*0.1025*
	Openness	50.06	49.88	0.5338	*0.0466*	*0.0511*

Table 6.2 Accuracy and correlation results of MBTI and Big Five based on the PANDORA dataset. Pear. and Spear. denote the Pearson/Spearman correlation between prediction and ground truth on each personality trait, *Italics* indicates statistical significance ($p < .05$).

6.2.2 ECM and Intent Predictor

Our empathetic signals comprise both ECM and intent, which are complementary. For example, *Encouraging* or *Sympathizing* in intent prediction is detailed beyond *Interpretation*

Traits	#Classes	Accuracy	Balanced_Accuracy	F1
ER	2	84.76	84.13	84.70
IP	2	84.12	85.35	84.23
EX	2	94.81	92.46	94.86
EI	9	90.17	90.17	90.23

Table 6.3 Evaluations on empathetic signals predictor. ER, IP, EX, and EI denote Emotional Reaction, Interpretation, Exploration, Empathetic Intent classification, respectively.

in the ECM. Additionally, ER within the ECM dictates whether a response contains emotional signals.

ECM: Inspired by Lee et al. [90], Fu et al. [91], Bi et al. [164], we use *IP*, *EX*, *ER* as parts of the empathetic signals. Specifically, *IP* represents expressions of acknowledgments or understanding of the interlocutor's emotion or situation. *EX* represents expressions of active interest in the interlocutor's situation; *ER* represents expressions of explicit emotions, as empathetic signals. Specifically, we follow official codes[3] and use three RoBERTa-based [50] classifiers to identify whether a response implies a certain trait individually.

Intent: Prior research by Welivita and Pu [1] highlighted incorporating dialogue intent modeling into response generation enhances the controllability and interpretability of generated responses. For this reason they introduced the EmpatheticIntents dataset,[4] which is enriched with intent annotations, such as *Questioning*, *Acknowledging*, and *Agreeing*. We fine-tune a RoBERTa-base [50] model on nine-class intent classification to label responses. The results are shown in Table 6.3.

6.3 Proposed Method

Fig. 6.2 shows an overview of our proposed method which comprises two main components. Firstly, a multi-grained prefix encoder is designed to implicitly learn the connections between personality traits and empathetic signals present in the system's response by multi-grained signals prediction and prefix encoding. Secondly, we introduce a personality enhancement mechanism aiming at integrating the generation of empathetic responses with explicit personality trait learning.

[3] https://github.com/behavioral-data/Empathy-Mental-Health

[4] https://github.com/anuradha1992/EmpatheticIntents

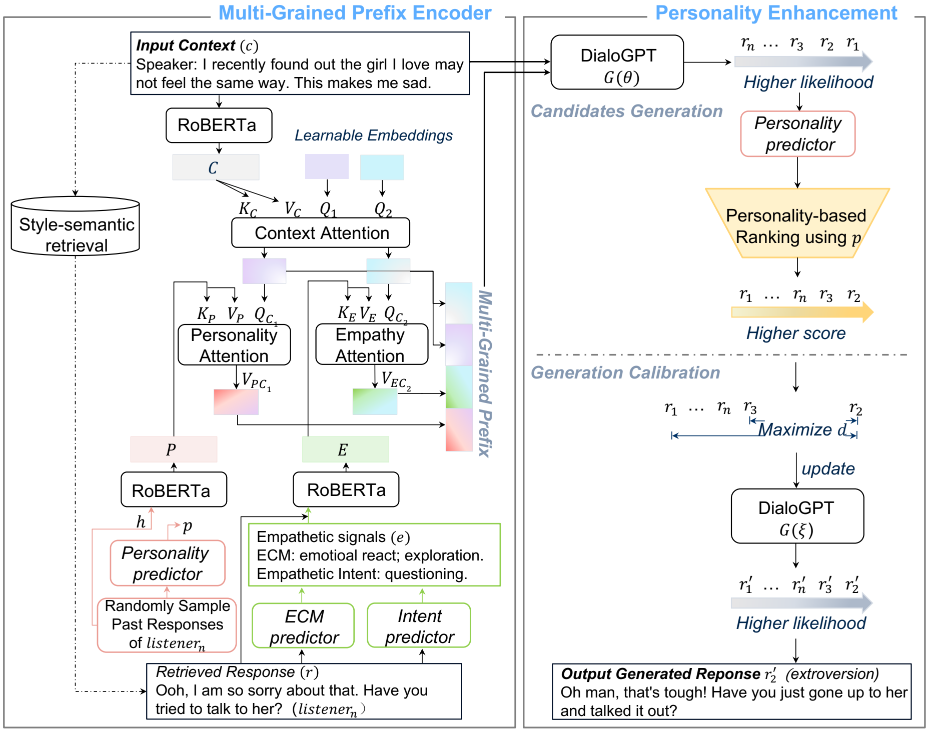


Fig. 6.2 The architecture of our proposed method that contains a multi-grained prefix encoder and personality enhancement module.

6.3.1 Mutli-Grained Prefix Encoder

There are 810 unique listeners in the benchmark ED dataset, and each participant is involved in up to 100 conversations. Based on the listener ID, we sampled past responses by the same listener from the training set to implicitly learn listener's personality. Inspired by the prefix-tuning mechanism employed in Li and Liang [32], Liu et al. [165], and Liu et al. [33], we project the input context (c), the concatenation of retrieved response (r) and empathy signals (e), and listener's past responses (h) into fixed-length prefix vectors, which are then prepended to the decoder hidden states as a prefix.

We first use the RoBERTa model to encode the c, e and h to continuous representations, denoted as $\mathbf{C}$, $\mathbf{P}$, $\mathbf{E}$:

$$\mathbf{C} = \text{RoBERTa}(c), \tag{6.1}$$

$$\mathbf{P} = \text{RoBERTa}(h), \tag{6.2}$$

$$\mathbf{E} = \text{RoBERTa}(\text{concat}(r, e)). \tag{6.3}$$

To separately extract distinct context-related empathy and personality features, we introduce two learnable embeddings to act as distinct queries, $\mathbf{Q_1}$ and $\mathbf{Q_2}$, where $\mathbf{Q_1}$ is in $\mathbb{R}^{dn_1}$ and $\mathbf{Q_2}$ in $\mathbb{R}^{dn_2}$; here, d represents the dimension of the RoBERTa's last hidden layer, while n_1 and n_2 denote the lengths of the respective queries. The context representation $\mathbf{C}$, serves as both key $\mathbf{K_C}$ and value $\mathbf{V_C}$. Employing a cross-attention mechanism, we project context $\mathbf{C}$ into two fixed-length prefix vectors. These vectors are subsequently treated as $\mathbf{Q_{C_1}}$ and $\mathbf{Q_{C_2}}$, respectively:

$$\mathbf{Q_{C_1}} = \mathrm{Attn}(\mathbf{K_C}, \mathbf{V_C}, \mathbf{Q_1}), \tag{6.4}$$

$$\mathbf{Q_{C_2}} = \mathrm{Attn}(\mathbf{K_C}, \mathbf{V_C}, \mathbf{Q_2}). \tag{6.5}$$

Then following the same process, we fuse the representations of the listener's past responses $\mathbf{P}$, and the empathy explanation representations $\mathbf{E}$, with the context-related prefix vectors $\mathbf{Q_{C_1}}$ and $\mathbf{Q_{C_2}}$, respectively:

$$\mathbf{V_{PC_1}} = \mathrm{Attn}(\mathbf{K_P}, \mathbf{V_P}, \mathbf{Q_{C_1}}), \tag{6.6}$$

$$\mathbf{V_{EC_2}} = \mathrm{Attn}(\mathbf{K_E}, \mathbf{V_E}, \mathbf{Q_{C_2}}). \tag{6.7}$$

This fusion process yields two distinct vectors: $\mathbf{V_{PC_1}}$, which encapsulates the context-personality relationship, and $\mathbf{V_{EC_2}}$, representing the context-empathy relationship. This ensures that both personality and empathy dimensions are considered in the context of the interaction.

We then concatenate $\mathbf{Q_{C_1}}$, $\mathbf{Q_{C_2}}$, $\mathbf{V_{PC_1}}$, and $\mathbf{V_{EC_2}}$ by the length dimension, followed by one linear layer, to produce the final representations $\mathbb{R}^{2(n_1+n_2)*d}$, as the final prefix embeddings.

6.3.2 Decoder

We utilize the pre-trained DialoGPT [52][5] as the decoder. We further feed the final prefix embeddings into DialoGPT-small and train the parameters in the model on the ED dataset, then obtain base empathetic response generator $G(\theta)$.

6.3.3 Personality Enhancement

Because the ED dataset primarily targets expressing empathy rather than personality, it is hard to learn personality traits from a single response with traditional backpropagation. Drawing inspiration from recent calibration work [166–168], we generate multiple candidate

[5] https://huggingface.co/docs/transformers/model-doc/dialogpt

responses via diverse beam search [169], which exhibit similar levels of empathy but vary in the degree of personality expressed. Subsequently, the proposed personality-based ranking module evaluates and ranks these candidates. Then, we calibrate the generation process by integrating a personality-oriented contrastive loss alongside the empathy loss, thereby achieving a generation of empathetic responses that reflect explicit personality traits.

Candidate Generation

For a input context c, we use the trained model $G(\theta)$ to generate K empathetic candidate responses by diverse beam search: $r_1, r_2, r_3, ..., r_K$, which can encapsulate varying degrees of personality expression.

Personality-based Ranking

We utilize our pre-trained personality predictor, which estimates the system's personality p from the past responses (h), including Big Five extroversion (p_e), MBTI introversion (p_i), and MBTI thinking (p_t). Then, we predict the personality traits of each candidate in $\{r_1, r_2, r_3, \ldots, r_K\}$, and calculate their personality margin S_{r_k}. This margin is derived as the sum of the mean square errors (MSE) between the personality scores p and the predicted scores for each trait, formulated as:

$$S_{r_k} = \left|p'_e - p_e\right|^2 + \left|p'_i - p_i\right|^2 + \left|p'_t - p_t\right|^2, \tag{6.8}$$

where p'_e, p'_i, and p'_t is the predicted score for each candidate on extroversion, introversion, and thinking traits, respectively. Following this, we re-rank all candidate responses based on the calculated personality margins in ascending order of S_{r_k}: $\{r'_1, r'_2, \ldots, r'_K\}$, where $S_{r'_i} < S_{r'_j}$, for $\forall i < j$.

Generation Calibration

We aim to encourage the model to assign higher estimated probabilities to empathetic candidate response with lower personality margin by adjusting the model $G(\theta)$ with a contrastive loss, following the previous work [166–168], the pairwise margin loss is defined as:

$$\mathscr{L}_p = \sum_i \sum_{j>i} \max(0, p(r'_j|c;\xi) - p(r'_i|c;\xi) + \lambda_{i,j}), \tag{6.9}$$

where $\lambda_{i,j}$ is the dynamic margin multiplied by the difference in rank between the candidates, $\lambda_{i,j} = \alpha * (j - i)$, and α is a hyper-parameter. $p(r_i^{'}|c;\xi)$ is the DialoGPT generation probability.

6.3.4 Training and Inference

Training During the training phase, we use the ground truth as the retrieved response for empathy and intent prediction. We aim to generate response that are both good at empathetic and personalized expression, then the final negative log-likelihood for generation is defined as:

$$\mathscr{L} = -\sum_{t=1}^{|y|} \log p\,(y_t|c, y_{<t};\xi) + \beta \mathscr{L}_p, \tag{6.10}$$

where β are hyper-parameters to balance empathy and personality loss. We minimize $\mathscr{L}$ to optimize the generator's parameters ξ.

Inference During the inference phase, we employ a style-semantic retrieval mechanism that matches each test set context (input) with similar contexts in the training set, selecting the most similar one's corresponding response as the retrieved response. Regarding the importance of emotion, semantics, and style in empathetic and personalized expression, we focus on these dimensions during the retrieval process. Specifically, we utilize Sentence-BERT [170][6] to obtain semantic embeddings. We employ an off-the-shelf, content-independent style representation model [171][7] for style embeddings. Furthermore, to enhance emotional relevance, we finetune RoBERTa [50][8] on the ED dataset, targeting a classification of 32 emotions, the accuracy of which is 56.06%. Subsequently, we extract emotional embeddings from the final layer of the fine-tuned RoBERTa model. The final retrieval score is:

$$\text{score} = \text{sim}_{sem} + \text{sim}_{style} + \text{sim}_{emo}, \tag{6.11}$$

where sim_{sem}, sim_{style}, and sim_{emo} represent similarity in semantics, style, and emotion, respectively.

6.4 Experimental Settings

We conducted experiments on the English EMPATHETICDIALOGUES dataset.

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[7] https://huggingface.co/AnnaWegmann/Style-Embedding

[8] https://huggingface.co/FacebookAI/roberta-base

6.4.1 Settings

Our implementation was based on Huggingface's Transformers.[9] For the multi-grained prefix encoder, we trained RoBERTa as an encoder and DialoGPT-small as the decoder from scratch on the ED dataset. We set the learning rate to 5e-5, and the batch size to 64. In the encoder configuration, the query length was set to 30. We sampled 10 past responses by the same listener from the training set. In the decoder configuration, the number of candidates K was set to 5. For the personality enhancement, we set α and β to be 0.001 and 1, separately. For the response generator, we used nucleus sampling (top-p) [172] with p set to 0.8 and temperature to 0.7. All experiments used the same seed to minimize the impact of randomness.

6.4.2 Models

Comparative Baselines

Transformer-based methods[10]:
MoEL [73]: which softly combines multiple emotion-specific decoders to a meta decoder to generate an empathetic response.
MIME [74]: which integrates emotion grouping, emotion mimicry, and stochasticity into the emotion mixture for various empathetic responses.
EmpDG [125]: which learns emotions and responses based on adversarial learning.
CEM [80]: which employs commonsense knowledge, to enhance its understanding of the interlocutor's situations and emotions.
Large language model-based methods:
DialoGPT [52]: a GPT2 model trained on Reddit conversation, we fine-tune it on the ED dataset for empathetic response generation.
LEMPEx [87]: which adopts T5 as the encoder-decoder and utilizes a combination of exemplar-based retrieval, a response generator, and an empathy control module to generate empathetic responses.[11]
ChatGPT+Causality [91]: which is based on a commonsense-based causality explanation that considers both the user's and the system's perspective to enhance ChatGPT's ability for empathetic response generation.

[9] https://huggingface.co/docs/transformers
[10] https://github.com/Sahandfer/CEM
[11] https://github.com/declare-lab/exemplary-empathy

Ablation Studies in Proposed StyEmp

We utilized DialoGPT as the base decoder across all ablation studies. The proposed StyEmp model integrates a multi-grained prefix encoder (MgPE (C+E+P)) with personality enhancement in the decoder (DialoGPT w/ PE). To explore the efficacy of each component within the encoder and decoder, we conducted ablation studies using four configurations of the multi-grained prefix encoder: (1) **MgPE (C+E+P)**: includes both the context-personality-aware prefix encoding and context-empathy-aware prefix encoding. In addition, there are other three configurations: (2) **MgPE (C)** incorporates only context-aware prefix encoding; (3) **MgPE (C+P)** includes only context-personality-aware prefix encoding; (4) **MgPE (C+E)** integrates only context-empathy-aware prefix encoding.

These were evaluated under two conditions in the decoder: **DialoGPT w/ PE** (with PE integration) and **DialoGPT w/o PE** (without PE integration).

6.4.3 Evaluation Metrics

Automatic Evaluations

BERTScore [124]: a BERT-based evaluation metric, which focuses on lexical semantic similarity between the generated response and the ground truth. We adopt its F1 score and use the "deberta-large-mnli" version.[12]

BLEURT [173]: evaluates to what extent the generated response is fluent and conveys the meaning of the reference.[13]

D1/D2 (Distinct-1/Distinct-2) [123]: calculates the number of distinct n-grams in generated responses.

E&I: denotes the mean Pearson correlation coefficient between the ground truth and generated responses for extroversion (E) from the Big Five predictor and introversion (I) from the MBTI predictor.

T: represents the Pearson correlation coefficient between the ground truth and generated responses for thinking (T) from the MBTI predictor.

EAcc.: refers to the average accuracy of both emotion (Emo.) and ER prediction, comparing the generated responses with ground truth.

IP&EX: refers to the average accuracy of both interpretation (IP) and exploration (EX) prediction, comparing generated responses with ground truth.

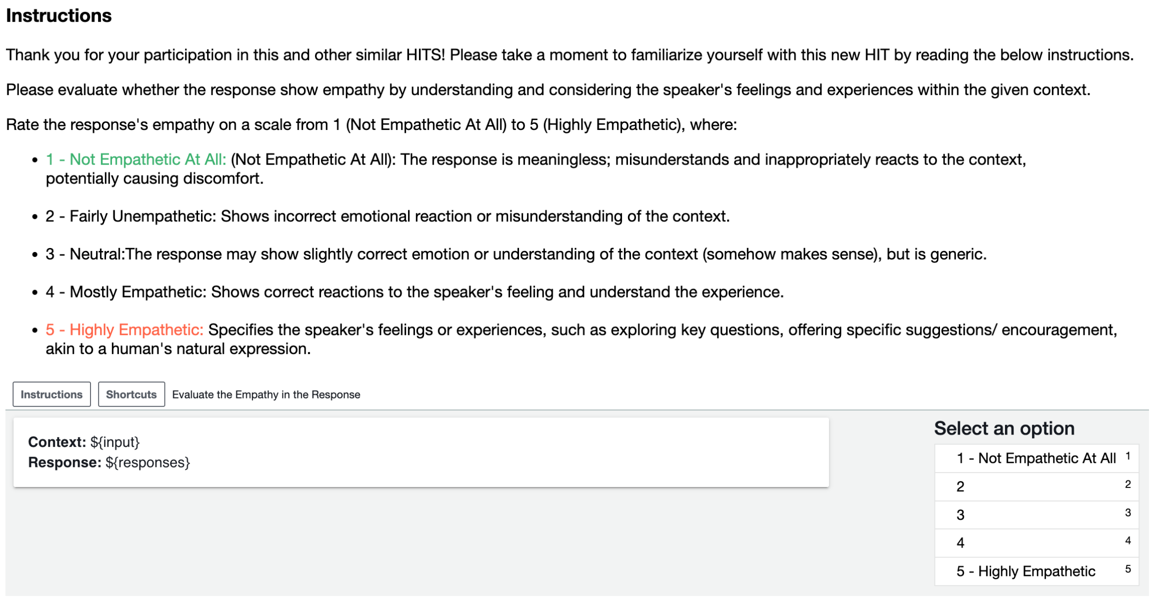
Intent: refers to the accuracy of empathetic intent prediction between the generated responses and ground truth.

[12] https://github.com/Tiiiger/bert_score

[13] https://github.com/google-research/bleurt

Human Evaluations

We randomly selected 100 context-response pairs from the test set across all models. Each response was evaluated by three different crowd-workers, provided with the corresponding context. We hired crowd workers through Amazon Mechanical Turk, and each has a historical approval rate of over 98% on human evaluation tasks. We assess the quality of these responses based on two criteria, each criterion is rated on a 1 to 5 scale: (1) **Empathy**, determining if the generated responses demonstrate understanding of the speaker's feelings and experiences. (2) **Personality**, refers to personality consistency; we provided crowd-workers with five sampled past responses from the listener of the ground truth and ask them to evaluate if the generated response aligns with the listener's personality traits. The template for the human evaluations is shown in Figs. 6.3 and 6.4.



Instructions

Thank you for your participation in this and other similar HITS! Please take a moment to familiarize yourself with this new HIT by reading the below instructions.

Please evaluate whether the response show empathy by understanding and considering the speaker's feelings and experiences within the given context.

Rate the response's empathy on a scale from 1 (Not Empathetic At All) to 5 (Highly Empathetic), where:

- 1 - Not Empathetic At All: (Not Empathetic At All): The response is meaningless; misunderstands and inappropriately reacts to the context, potentially causing discomfort.
- 2 - Fairly Unempathetic: Shows incorrect emotional reaction or misunderstanding of the context.
- 3 - Neutral:The response may show slightly correct emotion or understanding of the context (somehow makes sense), but is generic.
- 4 - Mostly Empathetic: Shows correct reactions to the speaker's feeling and understand the experience.
- 5 - Highly Empathetic: Specifies the speaker's feelings or experiences, such as exploring key questions, offering specific suggestions/ encouragement, akin to a human's natural expression.

Evaluate the Empathy in the Response

Context: ${input}
Response: ${responses}

Select an option

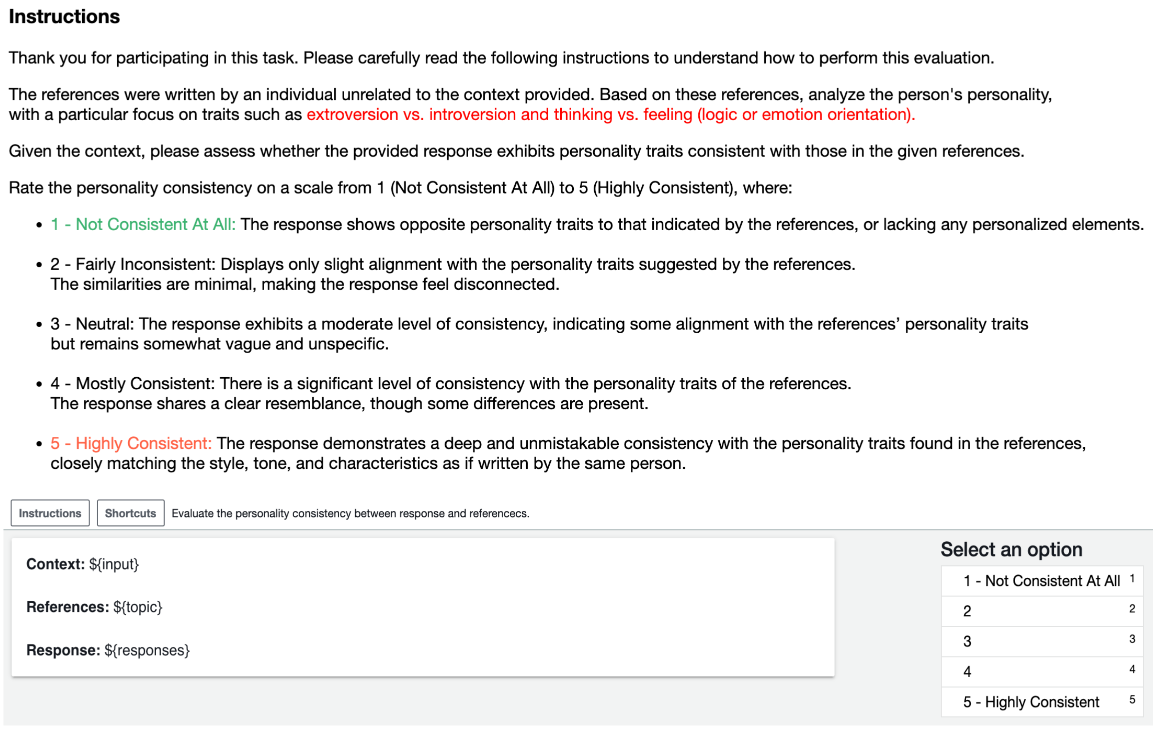
1 - Not Empathetic At All 1
2 2
3 3
4 4
5 - Highly Empathetic 5

Fig. 6.3 Template for human evaluation on empathy in generated responses.

6.5 Results and Analysis

6.5.1 Objective Evaluation Results

Table 6.4 presents the objective evaluation results for both comparative baselines (including Transformer-based and large language model-based methods), and our proposed method. The results illustrate that our method significantly outperforms the baselines in terms of personality, emotion, and intent accuracy, while maintaining the semantic scores comparable



Instructions

Thank you for participating in this task. Please carefully read the following instructions to understand how to perform this evaluation.

The references were written by an individual unrelated to the context provided. Based on these references, analyze the person's personality, with a particular focus on traits such as extroversion vs. introversion and thinking vs. feeling (logic or emotion orientation).

Given the context, please assess whether the provided response exhibits personality traits consistent with those in the given references.

Rate the personality consistency on a scale from 1 (Not Consistent At All) to 5 (Highly Consistent), where:

- 1 - Not Consistent At All: The response shows opposite personality traits to that indicated by the references, or lacking any personalized elements.
- 2 - Fairly Inconsistent: Displays only slight alignment with the personality traits suggested by the references. The similarities are minimal, making the response feel disconnected.
- 3 - Neutral: The response exhibits a moderate level of consistency, indicating some alignment with the references' personality traits but remains somewhat vague and unspecific.
- 4 - Mostly Consistent: There is a significant level of consistency with the personality traits of the references. The response shares a clear resemblance, though some differences are present.
- 5 - Highly Consistent: The response demonstrates a deep and unmistakable consistency with the personality traits found in the references, closely matching the style, tone, and characteristics as if written by the same person.

Evaluate the personality consistency between response and referencecs.

Context: ${input}

References: ${topic}

Response: ${responses}

Fig. 6.4 Template for human evaluation on personality consistency in generated responses.

to DialoGPT. The proposed StyEmp with PE degrades the semantic score because it re-ranks the original output of DialoGPT by weighting the personality consistency.

We also conducted ablation studies to evaluate different encoder configurations, comparing their performance in scenarios with and without PE. As depicted in Table 6.5, in both scenarios, MgPE (C+P) and MgPE (C+E) surpass MgPE (C) on most personality and empathy metrics. Moreover, MgPE (C+P+E) further outperforms both MgPE (C+P) and MgPE (C+E). These results substantiate our hypothesis that empathy and personality enrich each other. Incorporating PE further enhances the expression of both traits. These findings show the substantial contribution of the PE module in enhancing model performance for generating responses that are both empathetic and reflective of distinct personalities.

6.5.2 Human Evaluation Results

Table 6.6 shows that our methods rank highest against baselines. Specifically, DialoGPT with the proposed MgPE (C+E+P) and MgPE (C+E+P) w/ PE significantly outperform finetuned DialoGPT, enhancing empathy and personality expression in generated responses. However, StyEmp performs worse than MgPE (C+E) w/ PE and MgPE (C+E+P) w/o PE regarding personality, inconsistent with the objective evaluation results. This discrepancy stems from

Methods	Semantics		Diversity		Personality		Empathy		
	BERTS	BLEURT	D1	D2	E&I	T	EAcc.	IP&EX	Intent
Transformer-based methods									
MOEL	52.67	34.48	0.44	2.02	0.0525	0.0525	26.80	70.06	22.77
MIME	52.87	35.64	0.32	1.12	0.0200	0.0675	22.40	70.17	25.11
EmpDG	51.99	34.60	0.79	3.23	0.0155	0.1115	26.49	68.09	21.29
CEM	52.41	35.06	0.65	2.92	0.0741	0.1519	32.85	**73.62**	29.37
Large language model-based methods									
LEMPEx	49.03	27.92	1.20	12.88	-0.0077	0.0706	31.73	69.03	27.99
DialoGPT	54.24	40.32	2.92	15.62	0.1361	0.1723	33.68	72.49	31.53
ChatGPT+Causality	**54.93**	**43.45**	2.91	**16.44**	0.1584	0.1774	30.79	69.64	27.86
Our proposed method									
StyEmp w/o PE	54.13	41.00	**2.95**	16.10	0.1681	0.2010	34.47	72.70	31.73
StyEmp	53.60	40.49	2.21	9.48	**0.1758***	**0.2093***	**34.88***	73.02*	**31.85***

Table 6.4 Objective evaluation results of baselines and our proposed method. **Bold** and underline denote the best and second-best score, respectively. * indicates a statistically significant difference for $p < 0.05$ between StyEmp and ChatGPT+Causality, determined by t-test.

inaccuracies in personality prediction, particularly when conflicts arise between the predicted personality traits and those implied by past responses. This is a limitation of using personality predictor with accuracy of 60-70%. More error analysis can be found in Table 6.7.

6.5.3 Case Studies and Error Analysis

Table 6.8 compares our proposed StyEmp model with baseline methods, highlighting differences in personality trait expression. The baseline methods fall short of showing explicit personality traits, often resulting in more general responses. On the other hand, StyEmp showcases extroverted traits (predicted by our method), utilizing expressions like *"wow, bet"* and longer phrases. Moreover, the StyEmp-generated responses are more closely aligned with the personality traits shown in the ground truth, indicating its effectiveness in accurately reflecting personality.

We further show two examples that our StyEmp failed to show consistent personality because of incorrect personality prediction. In contrast, StyEmp without PE correctly expresses personality by learning from past responses by the same listener from the training set, as shown in Table 6.7.

Methods	Semantics		Diversity		Personality		Empathy		
	BERTS	BLEURT	D1	D2	E&I	T	EAcc.	IP&EX	Intent
DialoGPT w/o PE	54.24	40.32	2.92	15.62	0.1361	0.1723	33.68	72.49	31.53
+MgPE (C)	54.43	41.18	2.85	16.08	0.1525	0.1828	34.08	72.57	31.00
+MgPE (C+P)	53.99	40.31	**3.07**	**16.80**	0.1639	0.1987	34.30	71.71	31.47
+MgPE (C+E)	**54.55**	**41.25**	2.87	15.80	0.1552	0.1890	34.32	72.90	31.75
+MgPE (C+E+P)	54.13	41.00	2.95	16.10	0.1681	0.2010	34.47	72.70	31.73
DialoGPT w/ PE	53.92	40.37	2.23	9.74	0.1672	0.1824	34.37	73.42	**32.23**
+MgPE (C)	53.96	40.83	2.22	9.63	0.1669	0.1997	35.37	72.76	31.14
+MgPE (C+P)	53.24	40.29	2.05	8.93	0.1683	**0.2108**	34.14	72.81	31.42
+MgPE (C+E)	53.89	40.52	2.32	9.89	0.1680	0.1949	**35.65**	**73.58**	32.21
+MgPE (C+E+P)	53.60	40.49	2.21	9.48	**0.1758**	0.2093	34.88	73.02	31.85

Table 6.5 Ablation studies on the effect of context, past responses (implicit personality), empathy explanation in the multi-grained prefix encoder, and explicit personality enhancement.

Models	Empathy	Personality
CEM	3.35	2.93
ChatGPT+Causality	4.00	3.11
DialoGPT	3.04	2.99
DialoGPT+MgPE (C+E+P)	4.05*	3.25*
DialoGPT+MgPE (C+E) w/ PE	3.97	**3.39**
DialoGPT+MgPE (C+E+P) w/ PE	**4.08***	3.18*

Table 6.6 Results of human evaluations. DialoGPT+ MgPE (C+E+P) w/ PE refers to StyEmp. * indicates a statistically significant improvement ($p < 0.05$) over DialoGPT.

6.6 Conclusions and Future Work

We have proposed StyEmp, which aims to stylize the empathetic response generation with a consistent personality. Specifically, StyEmp incorporates a multi-grained prefix mechanism designed to capture the intricate relationship between a system's personality and its empathetic expressions. Furthermore, we introduce a personality enhancement module that leverages contrastive learning to calibrate the generation model, ensuring responses are both empathetic and reflective of a distinct personality. The experimental results demonstrate that our method outperforms other competitive methods on both automatic and human evaluations. The performance of our model is currently limited by the efficacy of the personality predictor. In future work, we plan to utilize ground truth personality traits instead of predicted ones by annotating the dataset with personality labels.

Context	I one time lost my trunks in the pool! People saw me in a way I didn't want!
Ground truth	Oh no! That must have been super embarrassing! How did you react to that?
Predicted system's	*personality: introvert, feeling*
Predicted system's	*Empathy: Emotional Reaction, Exploration. Emotion Intent is questioning.*
StyEmp w/o PE	Oh no! Did you get them back?
StyEmp	Did you get hurt?
Context	My friend came over yesterday and we were sitting on the couch chit-chatting and when I got up I accidentally farted. So embarrassing.
Ground truth	Oh my, did they notice you farted?
Predicted system's	*personality: introvert, feeling*
Predicted system's	*Empathy: Exploration. Emotion Intent is questioning.*
StyEmp w/o PE	Oh no! Did you say anything to him?
StyEmp	Did you apologize?

Table 6.7 Cases exist where StyEmp fails to accurately express the intended personality due to errors in personality prediction, which lead to errors in PE. In contrast, StyEmp without PE correctly expresses personality by learning from past responses by the same listener from the training set.

Limitations

Given our objective to enrich responses with empathy and personality information, we face the challenge of a scarcity of datasets that provide both empathy and personality annotations. Therefore, we have developed additional personality scorers, as shown in Table 6.1 and Table 6.2. However, the results from these scorers are not ideal, significantly impacting the effectiveness of our personality enhancement module, since we rely on the predicted personality to enhance the system's personality expression. To overcome this limitation, we plan to collect a dataset that includes both empathy and personality annotations in future work.

Context	I felt so nervous going to my last job interview! Nerves must occur in proportion to how badly you *want* something!
Ground truth	I know what you mean. Did you get the job?
MoEL	I am sorry to hear that. I hope you do well!
MIME	I am sure you will do well.
EmpDG	I am sure you will be fine.
CEM	I am sure you will be fine.
LEMPEx	That's a great job, I am sure you will do well.
DialoGPT	I'm sure you'll do great! I'm sure you'll do great!
ChatGPT+Causality	I understand how you feel, but remember to take deep breaths and focus on your strengths. You got this!
Predicted system's	*personality: introvert, feeling*
Predicted system's	*Empathy: Emotional reaction; Emotion intent is acknowledging.*
StyEmp w/o PE	I feel the same way. I hope you get the job!
StyEmp	I know what you mean. I'm nervous too.
Context	I went back to my home where I grew up and a flood of good memories came back to me. I had so many fun times with my best friends when I was a child.
Ground Truth	What is your favorite memory with your friends?
MoEL	That is so sad. I bet you were a great time.
MIME	That is so sweet. I am glad you had a good time.
EmpDG	I bet you were so proud of them.
CEM	That is awesome. I bet you were very happy.
LEMPEx	That must have been a cool old times!
DialoGPT	I bet you were so happy to have that feeling.
ChatGPT+Causality	That sounds amazing! I love reminiscing about my childhood too. Do you have any specific memories that stand out to you?
Predicted system's	*personality: extrovert, feeling*
Predicted system's	*Empathy: Interpretation. Emotion Intent is neutral.*
StyEmp w/o PE	I can imagine! It was nice to have good memories that you can cherish as a child. I'm glad you had good memories and not too many bad memories.
StyEmp	Oh wow, I bet that was a great time. I bet you were so happy to have them back.

Table 6.8 Comparative case studies between baselines and our proposed StyEmp.

Chapter 7

Conclusions

This chapter summarizes the contributions of this thesis, provides comparative case studies of our proposed methods, and outlines future research directions.

7.1 Contributions

This thesis addressed the empathetic response generation for spoken dialogue systems from the aspects of correct dialogue comprehension to dialogue personalization. To goal of this study is to ensure that the spoken dialogue system expresses the appropriate empathy and personality to the user to improve the human-system interaction experiences.

Dialogue Comprehension Correct dialogue comprehension is essential to generate an appropriate empathetic response. To this end, in Chapter 3, we have proposed a dual variational generative model for empathetic response generation. The proposed model can efficiently capture the mutual characteristics of the content and emotional consistency between the context and the response. Evaluations on both Japanese and English EMPATHETICDIALOGUES datasets demonstrate the proposed model's superiority in generating empathetic responses with contextual and emotional appropriateness. In addition to the DVG model, we proposed an auxiliary retrieval system to improve empathetic response generation. We further extended our model's potential to generate both empathetic and general responses and evaluated our system's effectiveness in enhancing human-robot interaction by a virtual agent. Subsequently, we integrated the system into a CommU humanoid robot for practical application.

In Chapter 4, we explored dialogue comprehension from the aspects of causality reasoning in the dialogue. We have proposed a commonsense-based causality explanation approach for diverse empathetic response generation that considers the system's intentions and reactions as well as the user's desires and reactions. Specifically, we enhance ChatGPT's ability to reason the system's intentions and reactions by integrating in-context learning with commonsense knowledge (desire, reaction, and intention). We have integrated the commonsense-based causality explanation with both ChatGPT and a trained T5 model. The experimental results demonstrate that our method outperforms other competitive methods on both automatic and human evaluations.

Dialogue Personalization Not only correct dialogue comprehension, but an appropriate empathetic response also depends on personality traits. Personality recognition in users and the development of systems that correspondingly express a consistent personality are crucial for enhancing the believability and engagement of human-robot interactions. To this end, in Chapter 5, we have proposed a data augmentation method for personality recognition, which involves interpolating between two existing data points to enhance speaker diversity. Additionally, we have introduced the HC-GNN method to independently model the interdependencies among interlocutors, as well as the intra-dependencies within the speaker in dialogues. Experimental results from the RealPersonaChat dataset demonstrate that increasing speaker diversity significantly improves personality recognition in both monologue and dialogue settings. Our HC-GNN method outperforms baseline models, showcasing its effectiveness.

In Chapter 6, we have proposed StyEmp, which aims to stylize empathetic response generation with consistent personality. Specifically, StyEmp incorporates a multi-grained prefix mechanism to capture the intricate relationship between a system's personality and its empathetic expressions. Furthermore, we introduce a personality reinforcement module that leverages contrastive learning to calibrate the generation model, ensuring responses are both empathetic and reflective of the distinct personality. The experimental results demonstrate that our method outperforms other competitive methods on both automatic and human evaluations.

7.2 Comparison of Proposed Methods

We provide an example to illustrate the comparison of our proposed methods and baselines for empathetic response generation, as shown in Table 7.1. The methods presented in Chapter 3 focus on generating appropriate empathetic responses that reflect both accurate affection and cognition by modeling the emotional and content consistency between the user's input and the system's output. Additionally, Chapter 3 incorporates a response retrieval model as a fallback to the generative model based on emotion recognition, leveraging the strengths of both generative and retrieval models. Chapter 4 further enhances the explainability

	Context	I studied so hard for 3 months straight for my bar exam to become a lawyer.
	Ground truth	wow, you're so determined! Did you pass your exam?
Baselines	MoEL	That is awesome! I hope you do well!
	MIME	That is great. I am sure you will do great!
	EmpDG	That is great! What did you do?
	CEM	that is great! I am sure you will do great!
	LEMPEx	Congratulations! That's awesome! Congratulations.
	DialoGPT	That's great, I hope you did well.
Chapter 3	DVG	That is awesome! I am sure you will do fine.
	Retrieval	I hope it goes well.
Chapter 4	*Predicted Causality$_{user}$*	*user wants: pass the exam; get a job.* *user reactions: proud; determined.*
	Reasoned Causality$_{sys}$	*sys intents: be successful; be proud.* *sys reactions: happy; accomplished.*
	T5+Causality$_{user,sys}$	That's awesome! I bet you were proud!
	ChatGPT+Causality$_{user,sys}$	That's amazing. I'm sure you will do great on the exam.
Chapter 6	*Predicted system's personality*	*extrovert, feeling*
	Predicted system's empathy	*emotional reaction; emotion intent is wishing.*
	StyEmp w/o PR	That's great! That's the best feeling in the world! What are you studying?
	StyEmp	Wow, that's a long time! I bet you were really proud of yourself! What kind of bar did you study? I hope you did well!

Table 7.1 Comparative case studies between baselines and our proposed methods.

and controllability of the empathetic response generation process by integrating causality reasoning from the user's want/reaction to the system's intent/reaction. In Chapter 6, we focus on stylizing the empathetic response generation to incorporate a suitable personality, which is closely related to distinct empathetic expressions and should be adapted to the user's personality. However, adapting the system's personality to best suit the user's personality will be addressed in future work.

7.3 Future Work

This section outlines several unresolved issues related to the methods developed in this thesis and proposes directions for future research.

User-adaptable Empathetic Dialogue Systems Humans with different personalities have varied preferences for systems personalities, highlighting the importance of user adaptation in dialogue systems to enhance user experience and engagement. To address this, Yamamoto et al. [174] introduced a method of user adaptation via character expression, where a dialogue system tailors its character to match the user's personality. Their model specifically utilized spoken features such as "Utterance Amount," "Backchannel," "Filler," and "Switching Pause," to control the system's character rather than the textual response. Moreover, appropriate empathetic expressions are also dependent on personality traits, making it crucial to adapt the system's empathetic style and personality to align with the user's personality in dialogues, thereby enhancing human-robot interactions.

Trustable evaluations for SDS In the field of open-domain dialogue systems, evaluation is commonly conducted using automatic metrics and human judgments. However, automatic metrics, such as BLEU, METEOR, and ROUGE, are based on word overlap and struggle to capture the diverse nature of dialogue systems. On the other hand, human judgments are more reliable, but expensive and lack standardized protocols. Hence, there exists a necessity to combine the merits of automated and human evaluations while mitigating their respective drawbacks. Inspired by Giorgi [175] who proposed human-centered metrics (such as emotion, and personality) for dialog system evaluation, hierarchical evaluation of spoken dialogue systems (SDS) represents a possible approach to effectively quantify system performance. For instance, at the utterance level within a conversation, to evaluate the "Relatedness," "Fluency," and "Informativeness" of the responses. Furthermore, at the conversation level, it is crucial to evaluate whether the responder demonstrates a distinct personality and exhibits empathy appropriately. Lastly, at the system level, the evaluation should consider the system's ability to maintain robustness across interactions with users possessing diverse personalities. However, since all the above evaluation aspects are subjective, the research of suitable automated metrics requires further exploration.

References

[1] Anuradha Welivita and Pearl Pu. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, 2020.

[2] Nadine R Richendoller and James B Weaver III. Exploring the links between personality and empathic response style. *Personality and individual Differences*, 17(3): 303–311, 1994.

[3] François Mairesse and Marilyn A Walker. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278, 2010.

[4] David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. Ada and grace: Direct interaction with museum visitors. In *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings 12*, pages 245–251. Springer, 2012.

[5] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. Job interviewer android with elaborate follow-up question generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 324–332, 2020.

[6] Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. An attentive listening system with android erica: Comparison of autonomous and woz interactions. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 118–127, 2020.

[7] Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. Erica: The erato intelligent conversational android. In *2016 25th IEEE International symposium on robot and human interactive communication (RO-MAN)*, pages 22–29. IEEE, 2016.

[8] G. Winata, H. Lovenia, E. Ishii, F. Siddique, Y. Yang, and P. Fung. Nora: The well-being coach. *ArXiv Preprint ArXiv:2106.00410*, 2021.

[9] G. Winata, O. Kampman, Y. Yang, A. Dey, and P. Fung. Nora the empathetic psychologist. In *INTERSPEECH*, pages 3437–3438, 2017.

[10] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623, 2020.

[11] M. Jung, Y. Lim, S. Kim, J. Jang, S. Shin, and K. Lee. An emotion-based korean multimodal empathetic dialogue system. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 16–22, 2022.

[12] Mark H Davis. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113, 1983.

[13] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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)*, pages 2204–2213, July 2018.

[14] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models. In *EMNLP*, pages 6556–6566, 2020.

[15] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[16] Isabel Briggs Myers. The myers-briggs type indicator: Manual (1962). 1962.

[17] Mostafa M Amin, Erik Cambria, and Björn W Schuller. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23, 2023.

[18] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, July 2019.

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[20] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, volume 33, pages 3027–3035, 2019.

[21] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.

[22] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[23] Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*, 2022.

[24] Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Shrivastava, Samson Tan, et al. Nl-augmenter: A framework for task-sensitive natural language augmentation. *arXiv preprint arXiv:2112.02721*, 2021.

[25] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pages 6382–6388, November 2019.

[26] Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, July 2020.

[27] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, August 2016.

[28] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[29] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[30] Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. Role play-based question-answering by real users for building chatbots with consistent personalities. In *19th SIGDIAL*, pages 264–272, 2018.

[31] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP*, 2021.

[32] 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)*, pages 4582–4597, 2021.

[33] Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. Recap: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, 2023.

[34] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[35] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. Artificial paranoia. *Artificial intelligence*, 2(1):1–25, 1971.

[36] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[37] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. Integration of speech recognition and natural language processing in the mit voyager system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pages 713–716. IEEE Computer Society, 1991.

[38] James Allen, Lenhart Schubert, George Ferguson, Peter Heeman, Chung Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, et al. The trains project: A case study in defining a conversational planning agent. 1994.

[39] George Ferguson, James F Allen, Bradford W Miller, et al. Trains-95: Towards a mixed-initiative planning assistant. In *AIPS*, pages 70–77, 1996.

[40] Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[41] Esther Levin, Roberto Pieraccini, and Wieland Eckert. Using markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 201–204. IEEE, 1998.

[42] Jason D Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[43] Milica Gasic, Filip Jurcicek, Simon Keizer, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Proceedings of the SIGDIAL 2010 Conference*, pages 201–204, 2010.

[44] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, 2018.

[45] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[46] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

[50] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[51] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[52] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.

[53] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21 (1):5485–5551, 2020.

[54] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[56] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[57] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[58] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[60] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381, 2019.

[61] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[63] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018.

[64] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[65] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

[66] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

[67] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, 2020.

[68] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

[69] Laurel D Riek, Philip C Paul, and Peter Robinson. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces*, 3:99–108, 2010.

[70] Frank Hegel, Torsten Spexard, Britta Wrede, Gernot Horstmann, and Thurid Vogt. Playing a different imitation game: Interaction with an empathic android robot. In *2006 6th IEEE-RAS international conference on humanoid robots*, pages 56–61. IEEE, 2006.

[71] Doori Jo, Jooyun Han, Kyungmi Chung, and Sukhan Lee. Empathy between human and robot? In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 151–152. IEEE, 2013.

[72] Koji Inoue, Divesh Lala, and Tatsuya Kawahara. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9: 933261, 2022.

[73] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, pages 121–132, 2019.

[74] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*, pages 8968–8979, 2020.

[75] Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. E-core: Emotion correlation enhanced empathetic dialogue generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[76] Hyunwoo Kim, Byeongchang Kim, and Gun Hee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2227–2240. Association for Computational Linguistics (ACL), 2021.

[77] Jiashuo Wang, Yi Cheng, and Wenjie Li. Care: Causality reasoning for empathetic responses by conditional graph generation. *EMNLP findings*, 2022.

[78] Zi Haur Pang, Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. Acknowledgment of emotional states: Generating validating responses for empathetic dialogue. *Proc. Int’ l Workshop Spoken Dialogue Systems (IWSDS)*, 2024.

[79] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. Knowledge bridging for empathetic dialogue generation. 2022.

[80] Sahand Sabour, Chujie Zheng, and Minlie Huang. Cem: Commonsense-aware empathetic response generation. In *AAAI*, volume 36, pages 11229–11237, 2022.

[81] Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot. *Advanced Robotics*, 37(21):1406–1418, 2023.

[82] Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. Improving empathetic response generation with retrieval based on emotion recognition. 2023.

[83] Zhongyang Li, Xiao Ding, Ting Liu, J Edward Hu, and Benjamin Van Durme. Guided generation of cause and effect. *arXiv preprint arXiv:2107.09846*, 2021.

[84] Sahand Sabour, Chujie Zheng, and Minlie Huang. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*, 2021.

[85] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, 2020.

[86] Zhou Yang, Zhaochun Ren, Yufeng Wang, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Yunbing Wu, Yisong Su, Sibo Ju, and Xiangwen Liao. Exploiting emotion-semantic correlations for empathetic response generation. *arXiv preprint arXiv:2402.17437*, 2024.

[87] Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10: 77176–77190, 2022.

[88] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[89] Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*, 2021.

[90] Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *29th COLING*, pages 669–683, 2022.

[91] Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response generation. In *24th SIGDIAL*, pages 645–656, 2023.

[92] Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, 2024.

[93] Zhengjie Huang, Pingsheng Liu, Gerard de Melo, Liang He, and Linlin Wang. Generating persona-aware empathetic responses with retrieval-augmented prompt learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12441–12445. IEEE, 2024.

[94] Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. Pecer: Empathetic response generation via dynamic personality extraction and contextual emotional reasoning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10631–10635. IEEE, 2024.

[95] Yahui Fu, Chenhui Chu, and Tatsuya Kawahara. Styemp: Stylizing empathetic response generation via multi-grained prefix encoder and personality reinforcement. *arXiv preprint arXiv:2408.02271*, 2024.

[96] Ling Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. Multi-party empathetic dialogue generation: A new task for dialog systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–307, 2022.

[97] Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. Multimodal learning for identifying opportunities for empathetic responses. In *2019 International Conference on Multimodal Interaction*, pages 95–104, 2019.

[98] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. Empathyear: An open-source avatar multimodal empathetic chatbot. *arXiv preprint arXiv:2406.15177*, 2024.

[99] Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Won Park. Empathicstories++: A multimodal dataset for empathy towards personal experiences. *Findings of ACL*, 2024.

[100] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based Japanese chit-chat systems. *arXiv preprint arXiv:2109.05217*, 2021.

[101] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors’ own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 2023.

[102] D. McNeill and C. Kennington. Predicting human interpretations of affect and valence in a social robot. In *Proceedings of Robotics: Science and Systems*, 2019.

[103] A. Lee and H. Ishiguro. Development of cg-based embodied dialogue agents and system with conversational reality for avatar-symbiotic research. *SIG-SLUD, JSAI*, 2022.

[104] Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3124–3134, 2021.

[105] Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. A generative model for joint natural language understanding and generation. In *ACL*, 2020.

[106] Shaobo Cui, Rongzhong Lian, Yuanfeng Song Di Jiang, Siqi Bao, and Yong Jiang. DAL: Dual adversarial learning for dialogue generation. In *NAACL Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 11–20, 2019.

[107] Xuemeng Hu, Rui Wang, Deyu Zhou, and Yuxuan Xiong. Neural topic modeling with cycle-consistent adversarial training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9018–9030, 2020.

[108] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, 2019.

[109] Yizhe Zhang, Siqi Sun, Xiang Gao, Yuwei Fang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Retgen: A joint framework for retrieval and grounded text generation modeling. 2022.

[110] Hao Shi, Longbiao Wang, Meng Ge, Sheng Li, and Jianwu Dang. Spectrograms fusion with minimum difference masks estimation for monaural speech dereverberation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7544–7548, 2020. doi: 10.1109/ICASSP40776.2020.9054661.

[111] Hao Shi, Longbiao Wang, Sheng Li, Chenchen Ding, Meng Ge, Nan Li, Jianwu Dang, and Hiroshi Seki. Singing Voice Extraction with Attention-Based Spectrograms Fusion. In *Proc. Interspeech 2020*, pages 2412–2416, 2020. doi: 10.21437/Interspeech.2020-1043.

[112] Lili Guo, Longbiao Wang, Jianwu Dang, Yahui Fu, Jiaxing Liu, and Shifei Ding. Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information. *IEEE MultiMedia*, 29(2):94–103, 2022.

[113] Hao Shi, Masato Mimura, and Tatsuya Kawahara. Waveform-domain speech enhancement using spectrogram encoding for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3049–3060, 2024. doi: 10.1109/TASLP.2024.3407511.

[114] Hao Shi, Longbiao Wang, Sheng Li, Cunhang Fan, Jianwu Dang, and Tatsuya Kawahara. Spectrograms fusion-based end-to-end robust automatic speech recognition. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 438–442, 2021.

[115] Hao Shi and Tatsuya Kawahara. Dual-path Adaptation of Pretrained Feature Extraction Module for Robust Automatic Speech Recognition. In *Proc. Interspeech 2024*, 2022. doi: 10.21437/Interspeech.2022-11423.

[116] Hao Shi and Tatsuya Kawahara. Exploration of adapter for noise robust automatic speech recognition. *arXiv preprint arXiv:2402.18275*, 2024.

[117] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[118] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[119] Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, pages 10–21, 2016.

[120] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[121] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[122] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. In *EMNLP*, pages 1442–1451, 2017.

[123] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, pages 110–119, 2016.

[124] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

[125] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. Empdg: Multi-resolution interactive empathetic dialogue generation. In *28th COLING*, pages 4454–4466, 2020.

[126] Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715*, 2022.

[127] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, volume 35, pages 6384–6392, 2021.

[128] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[129] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 11 2019. URL http://arxiv.org/abs/1908.10084.

[130] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[131] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500, nov 2007. ISSN 1076-9757.

[132] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[133] Majid Ramezani, Mohammad-Reza Feizi-Derakhshi, and Mohammad-Ali Balafar. Text-based automatic personality prediction using kgrat-net: a knowledge graph attention network classifier. *Scientific Reports*, 12(1):21453, 2022.

[134] Zhiyuan Wen, Jiannong Cao, Yu Yang, Haoli Wang, Ruosong Yang, and Shuaiqi Liu. Desprompt: Personality-descriptive prompt tuning for few-shot personality recognition. *Information Processing & Management*, 60(5):103422, 2023.

[135] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 400–418. Springer, 2016.

[136] M. M. Amin, E. Cambria, and B. W. Schuller. Can chatgpt' s responses boost traditional natural language processing? *IEEE Intelligent Systems*, 38(05):5–11, sep 2023. ISSN 1941-1294. doi: 10.1109/MIS.2023.3305861.

[137] Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, June 2021.

[138] Esteban Andres Rissola, Seyed Ali Bahrainian, and Fabio Crestani. Personality recognition in conversations using capsule neural networks. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 180–187, 2019.

[139] Hang Jiang, Xianzhe Zhang, and Jinho D Choi. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13821–13822, 2020.

[140] Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. CPED: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*, 2022. URL https://arxiv.org/abs/2205.14727.

[141] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862, 2006.

[142] Wenjing Han, Yirong Chen, Xiaofen Xing, Guohua Zhou, and Xiangmin Xu. Speaker-aware hierarchical transformer for personality recognition in multiparty dialogues. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[143] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[144] Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, July 2020.

[145] Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. A task in a suit and a tie: paraphrase generation with semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7176–7183, 2019.

[146] Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, June 2018.

[147] Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, August 2019.

[148] Maxwell Mojapelo and Jan Buys. Data augmentation for low resource neural machine translation for sotho-tswana languages. In *Proceedings of Southern African Conference for AI Research (SACAIR 2023)*, Johannesburg, South Africa, 2023.

[149] Jason Wei. Good-enough example extrapolation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5923–5929, November 2021.

[150] Yahui Fu, Lili Guo, Longbiao Wang, Zhilei Liu, Jiaxing Liu, and Jianwu Dang. A sentiment similarity-oriented attention model with multi-task learning for text-based emotion recognition. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part I 27*, pages 278–289. Springer, 2021.

[151] Hao Shi, Kazuki Shimada, Masato Hirano, Takashi Shibuya, Yuichiro Koyama, Zhi Zhong, Shusuke Takahashi, Tatsuya Kawahara, and Yuki Mitsufuji. Diffusion-based speech enhancement with joint generative and predictive decoders. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12951–12955, 2024. doi: 10.1109/ICASSP48485.2024.10448429.

[152] Hao Shi, Masato Mimura, Longbiao Wang, Jianwu Dang, and Tatsuya Kawahara. Time-domain speech enhancement assisted by multi-resolution frequency encoder and decoder. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094718.

[153] Hao Shi, Longbiao Wang, Sheng Li, Jianwu Dang, and Tatsuya Kawahara. Monaural Speech Enhancement Based on Spectrogram Decomposition for Convolutional Neural Network-sensitive Feature Extraction. In *Proc. Interspeech 2022*, pages 221–225, 2022. doi: 10.21437/Interspeech.2022-11268.

[154] Hao Shi, Yuchun Shu, Longbiao Wang, Jianwu Dang, and Tatsuya Kawahara. Fusing multiple bandwidth spectrograms for improving speech enhancement. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1938–1943, 2022. doi: 10.23919/APSIPAASC55919.2022.9980180.

[155] Jiaxing Liu, Sen Chen, Longbiao Wang, Zhilei Liu, Yahui Fu, Lili Guo, and Jianwu Dang. Multimodal emotion recognition with capsule graph convolutional based

representation fusion. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6339–6343. IEEE, 2021.

[156] Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. CONSK-GCN: conversational semantic-and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[157] Yahui Fu, Okada Shogo, Longbiao Wang, Guo Lili, Song Yaodong, Liu Jiaxing, and Dang Jianwu. Context-and knowledge-aware graph convolutional network for multimodal emotion recognition. *IEEE MultiMedia*, 29(3):91–100, 2022.

[158] Fu Changzeng, Chen Zhenghan, Shi Jiaqi, Wu Bowen, Liu Chaoran, Ishi Carlos Toshinori, and Ishiguro Hiroshi. Hag: Hierarchical attention with graph network for dialogue act classification in conversation. In *ICASSP*, pages 1–5. IEEE, 2023.

[159] Yang Tao, Deng Jinghao, Quan Xiaojun, and Wang Qifan. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904, 2023.

[160] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[161] Yamada Ikuya, Asai Akari, Shindo Hiroyuki, Takeda Hideaki, and Matsumoto Yuji. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020.

[162] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[163] Gordon Willard Allport. Personality: A psychological interpretation. 1937.

[164] Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin, and Xiaodong He. Diffusemp: A diffusion model-based framework with multi-grained control for empathetic response generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[165] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022.

[166] Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*, 2022.

[167] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, 2022.

[168] WANG Jiashuo, Haozhao Wang, Shichao Sun, and Wenjie Li. Aligning language models with human preferences via a bayesian approach. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[169] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[170] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992, 2019.

[171] Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, 2022.

[172] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

[173] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 2020.

[174] Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. Character expression of a conversational robot for adapting to user personality. *Advanced Robotics*, 38(4): 256–266, 2024.

[175] Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H Ungar, H Andrew Schwartz, and Joao Sedoc. Human-centered metrics for dialog system evaluation. *arXiv preprint arXiv:2305.14757*, 2023.

List of Publications

Journal Articles

1. Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. "Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot." *Advanced Robotics* 37, no. 21 (2023): 1406-1418. (**Chapter 3**)

International Conferences

1. Yahui Fu, Chenhui Chu, and Tatsuya Kawahara. "StyEmp: Stylizing Empathetic Response Generation via Multi-Grained Prefix Encoder and Personality Reinforcement." In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, 2024. (accepted) (**Chapter 6**)

2. Yahui Fu, Haiyue Song, Tianyu Zhao, and Tatsuya Kawahara. "Enhancing Personality Recognition in Dialogue by Data Augmentation and Heterogeneous Conversational Graph Networks." In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2024. (**Chapter 5**)

3. Zi Haur Pang , Yahui Fu, Divesh Lala, Keiko Ochi, Koji Inoue, and Tatsuya Kawahara. "Acknowledgment of Emotional States: Generating Validating Responses for Empathetic Dialogue." In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2024.

4. Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. "Reasoning before Responding: Integrating Commonsense-based Causality Explanation for Empathetic Response Generation." In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 645-656, 2023. (**Chapter 4**)

5. Yahui Fu, Koji Inoue, Divesh Lala, Kenta Yamamoto, Chenhui Chu, and Tatsuya Kawahara. "Improving Empathetic Response Generation with Retrieval based on Emotion Recognition." In *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2023. (**Chapter 3**)