인지과학, 제35권 제1호
Korean Journal of Cognitive Science
2024, Vol. 35, No. 1, 23～48
http://dx.doi.org/10.19066/cogsci.2024.35.1.002

Enhancing Empathic Reasoning of Large Language Models Based on Psychotherapy Models for AI-assisted Social Support

Yoon Kyung Lee[1)†] Inju Lee[1)] Minjung Shin[2)] Seoyeon Bae[1)] Sowon Hahn[1,2)†]

[1)]Department of Psychology, Seoul National University

[2)]Interdisciplinary Program in Cognitive Science, Seoul National University

Building human-aligned artificial intelligence (AI) for social support remains challenging despite the advancement of Large Language Models. We present a novel method, the Chain of Empathy (CoE) prompting, that utilizes insights from psychotherapy to induce LLMs to reason about human emotional states. This method is inspired by various psychotherapy approaches－Cognitive-Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), Person-Centered Therapy (PCT), and Reality Therapy (RT)－each leading to different patterns of interpreting clients' mental states. LLMs without CoE reasoning generated predominantly exploratory responses. However, when LLMs used CoE reasoning, we found a more comprehensive range of empathic responses aligned with each psychotherapy model's different reasoning patterns. For empathic expression classification, the CBT-based CoE resulted in the most balanced classification of empathic expression labels and the text generation of empathic responses. However, regarding emotion reasoning, other approaches like DBT and PCT showed higher performance in emotion reaction classification. We further conducted qualitative analysis and alignment scoring of each prompt-generated output. The findings underscore the importance of understanding the emotional context and how it affects human-AI communication. Our research contributes to understanding how psychotherapy models can be incorporated into LLMs, facilitating the development of context-aware, safe, and empathically responsive AI.

Key words : Empathy, AI-assisted Social Support, Large Language Models, Natural Language Processing, Cognitive AI, AI Alignment

† 교신저자: 이윤경, 한소원, 서울대학교 심리학과 인간공학심리 연구실, (00826) 서울특별시 관악구 관악로 1 16동 M413호
연구분야: 인지심리, 정서, 사회인지, 인간-인공지능 상호작용, 자연어처리
E-mail: yoonlee78@snu.ac.kr, swhahn@snu.ac.kr

Introduction

The quest to discern whether machines can possess 'mind' and exhibit human-level abilities has been a persistent endeavor, bridging discipline such as psychology, linguistics, and artificial intelligence (Turing, 1950). Recent advances in Large Language Models (LLMs), fueled by extensive pre-trained dataset from vast web corpora, have led to significant enhancements in language generation capabilities (Brown et al., 2020). These models exhibit proficiency in generating text that often mirrors human expressions, although the extent and nature of this resemblance continue to be a subject of scientific investigation (Bommasani et al., 2021; Brown et al., 2020; Taori et al., 2023; Touvron and Lavril et al., 2023). These models have also demonstrated effectiveness in problem-solving tasks, as evidenced by their high-performance metrics in contexts like professional exams, including the bar exam (Bommarito II and Katz, 2022), mathematical assessments (Zhang et al., 2023), and medical diagnostic scenarios (Nori et al., 2023).

As the scale of LLMs expands, a variety of capabilities and phenomena are emerging, including the potential for human-level cognitive abilities. It has been recently discovered that these models can be prompted to reason in a manner similar to humans, enhancing their performance in areas previously limited, such as mathematical logic and reasoning. A notable development in this context is the introduction of *Chain-of-Thought* (CoT) prompting (Kojima et al., 2022; Wei et al., 2022), which has proven to be instrumental in improving model performance by explicating the thought process leading to the conclusions. Unlike previous models, which provided answers without revealing the underlying reasoning process, CoT enables the explanation and interpretation of the reasoning behind the results. Nevertheless, this recent method has primarily experimented with logical or arithmetic tasks. Whether reasoning about emotional states or underlying causes enhances empathic responses to user input remains a relatively under-explored area and merits investigation.

Empathic understanding in humans involves cognitively reasoning about others' mental states. Recognizing that empathic responses are complex cognitive processes, not just emotional reaction, highlights the potential of integrating empathy into LLMs. By incorporating these human cognitive processes, LLMs can offer more nuanced and contextually relevant empathic responses.

Various psychotherapy models offer insights into empathic interactions, emphasizing the importance of accurately inferring emotions and situational factors for better social support and emotion regulation (Cooper & McLeod, 2011; Hofmann et al., 2010; Linehan, 1987; Wubbolding et al., 2017).

Incorporating these insights into LLMs' reasoning processes can enhance their empathic response accuracy and alignment with users' needs.

For this purpose, this study explores these possibilities and proposes a novel approach: *Chain-of-Empathy* (CoE) prompting. The CoE prompt integrates a reasoning of emotions and the specific factors contributing to these emotions, such as cognitive errors or unmet needs, before generating the response to users' input. In addition, we investigated whether these reasoning steps and responses align with the objectives of each psychotherapy model. This was done by conducting both automatic and qualitative evaluations using an existing benchmark dataset for empathic responses.

Related Work

Multi-construct of Empathy

Empathy, defined as the sharing of others' emotions and experiences, is a multifaceted concept encompassing cognitive, emotional, and motivational aspects (Anderson & Keltner, 2002; De Vignemont & Singer, 2006; Hall & Schwartz, 2019; Neff, 2003; Zaki, 2019). Cognitive empathy involves understanding others' emotions and perspectives linked to abilities such as mentalizing. For example, we use cognitive empathy when discerning other people's mental states by differentiating our own and others' mental states (Davis, 1980; 1983; Eisenberg, 2014). This form of empathy necessitates a comprehensive cognitive appraisal of situations, considering factors such as pleasantness, control, and certainty of outcomes, akin to a negotiator assessing both sides of a conflict (Lazarus, 1991; Wondra & Ellsworth, 2015). Affective empathy enables individuals to vicariously experience others' emotions, as seen when a person feels joy watching a friend achieve a goal or showing empathic concern for someone in a more unfortunate situation than self (Davis, 1980; 1983). Motivational empathy, a more recent facet of the multi-construct theory of empathy in social psychology, involves the drive to alleviate others' emotional distress, often leading to an altruistic attitude and prosocial behavior. For example, this can be observed in a bystander's compulsion to assist someone in an unfortunate situation, such as signing a petition for more affordable housing plans to help people who are homeless (Herrera et al., 2018; Zaki, 2019).

AI-assisted Social Support

Natural Language Processing (NLP) has been increasingly employed in developing conversational agents across various professional domains. Past research mainly targeted specific domains like social media, online peer support platforms, or mobile messenger, where much research focuses on advising users to share their concerns or extending condolences during adverse events (e.g., Ta et al., 2020; Wang et al., 2021). Social support extends to domains like psychotherapy, which conventionally have taken only in-person interactions (e.g., Fitzpatrick et al., 2017; Inkster et al., 2018; Mehta et al., 2021). However, with the increasing availability of online platforms, researchers now access large-scale data capturing people's expressions of emotions and daily lives, exemplified by open-source datasets like EmpatheticDialogues (Rashkin et al., 2018) and EPITOME (Sharma et al., 2020). These resources are instrumental in enhancing the communication abilities of conversational agents, particularly in properly reflecting human cognition and emotion (e.g., Casas et al., 2021; Gruschka et al., 2023; Xie & Park, 2021).

Historically, psychotherapy research has focused on empathy through nonverbal cues like body language and facial expressions, often necessitating manual coding of empathic responses (American Psychiatric Association, 1994; Ekman and Friesen, 1971; Scherer et al., 2001). However, AI advancements have led to a computational approach, predicting empathy from text corpora and quantifying it by labeling emotions (Rashkin et al., 2019) and distress (Buechel et al., 2018). While past studies mostly focused on the client's capacity for empathy, the counselor's expressed empathy is increasingly vital for successful therapy outcomes (Truax and Carkhuff, 2007). This aspect of expressed empathy aligns with our approach to utilizing LLMs to reflect the client's needs accurately.

Psychotherapy techniques have been applied to building mental health chatbots to enhance user engagement and therapeutic outcomes. Cognitive-Behavioral Therapy (CBT), established by Beck (1976), remains the most widely used approach both human-human and human-chatbot interactions because of its systematic nature (Fitzpatrick et al., 2017; Inkster et al., 2018). An example of an application that uses CBT model to assist its users is Woebot. This chatbot has been shown to have significant potential in supporting long-term treatment for anxiety and depression, as demonstrated by recent studies (Fitzpatrick et al., 2017; Nwosu et al., 2022).

Beyond CBT, other psychotherapy models such as Dialectical Behavior Therapy (DBT), Person-Centered Therapy (PCT), and Reality Therapy (RT) have been adopted in conversational agents to cater to a broader range of mental health issues. DBT, an extension of CBT, is a type of therapy that focuses on improving emotion regulation and interpersonal challenges. It has been used for individuals with borderline personality disorder (Linehan, 1987; 1991). Wysa, a mobile chatbot app, integrates CBT and DBT approaches to better understand and respond to users' emotional states (Inkster et al., 2018). PCT, pioneered by Rogers in 1957, emphasizes empathy, unconditional positive regard, and rapport-building in therapy. This method enables the development of chatbots like Serena (Brocki et al., 2023), which uses deep learning to create a dialogue system based on the PCT model. RT, established by Glasser in 1965, takes a practical, action-oriented approach that focuses on addressing individuals' current issues and finding solutions. Within RT, the 'WDEP framework (Wants, Direction, Evaluation, and Planning)' provides a clear structure for self-evaluation. This framework assists clients in reflecting on and strategizing their present abilities without dwelling on the past (Wubbolding et al., 2017).

By incorporating varied approaches from psychology, conversational agents can provide more personalized and effective support, reflecting users' diverse needs and preferences; This involves providing effective support for victims of crime (Ahn et al., 2020), individuals with autism (Diehl et al., 2012), and those with anxiety disorders (Rasouli et al., 2022). To enhance the empathetic communication abilities of conversational agents, it is imperative to adopt a multidisciplinary and specifically computational approach.

AI Empathy

The concept of AI's empathy, while lacking a consensus in definition, is generally regarded as the ability of AI to have similar reactions toward one's emotional experience as human's empathic reactions (Concannon & Tomalin, 2023; Paiva et al., 2017). This is explored through two primary approaches (Concannon & Tomalin, 2023). The first approach emphasizes a foundational similarity, suggesting that AIs should undergo developmental processes like humans' developmental stages through interaction with their environment and physical embodiment (Asada, 2015). This perspective posits that AI should mirror the human developmental trajectory in developing social intelligence.

In contrast, the second approach replicates human empathic expressions, sidelining the internal mechanisms that drive such responses. In this approach, AI's empathic abilities are viewed as the capacity to appropriately mimic human responses, particularly in verbal communication (e.g., Casas et al., 2021; Rashkin et al., 2019; Welivita et al., 2021). Currently, most AI empathy research aligns with this second approach, focusing on the outward manifestation of empathy rather than embedding a human-like developmental process into the AI system (Concannon & Tomalin, 2023).

Human assessment of AI's empathic abilities involves evaluating AI responses using various criteria. These evaluations typically break down empathy into several sub-domains. For example, Paiva et al. (2017) developed a questionnaire focusing on empathic concern and perspective-taking. Charrier et al. (2019) introduced the Robot's Perceived Empathy (RoPE) scale, distinguishing between empathic understanding and response. Casas et al. (2021) had human judges assess chatbot empathy, covering emotion reflection and well-being impact. Similarly, Lee and Yi (2023) proposed the Agent Empathic Reactivity Index (AERI), which measures empathy across four dimensions: perspective-taking, fantasy, empathic concern, and personal distress. Lastly, the Empathy Scale for Human-Computer Communication (ESHCC) by Concannon and Tomalin (2023) evaluates perceived empathy in AI-human interactions, considering affective, cognitive, attitudinal, and attunement factors.

The computational approach to AI empathy involves measuring how closely responses generated by AI match with the established standards of human empathy, also referred to as alignment between AI and human standards (Gabriel, 2020). These methods are mostly developed in interdisciplinary domains like Human-Computer Interaction (HCI), Natural Language Processing (NLP), Computational Linguistics, and Computer Science. For instance, Rashkin et al. (2019) employed Bilingual Evaluation Understudy (BLEU) scores to gauge the degree of congruence between AI-generated responses and typical human replies. Similarly, Zhou et al. (2020) analyzed indicators such as the frequency of message exchanges and the level of user engagement to assess the empathic effectiveness of AI. These advanced models offer a promising avenue for refining and better articulating the mechanisms behind AI's empathic responses, representing a significant leap in our ability to design AI systems that more closely mirror human empathic understanding.

Aligning Large Language Models for Enhanced Empathic Human-AI Interaction

LLMs are pre-trained transformer-based generative models. These models, a subset of artificial intelligence algorithms, are capable of generating realistic text responses. The latest language model advancements include OpenAI's Generative Pre-trained Transformer (GPT) series – GPT-3, GPT-3.5, and GPT-4 (Brown et al., 2020), as well as Meta AI's Large Language Model (LLaMA) series -

LLaMa and LLaMa 2-Chat (Touvron et al., 2023a; 2023b). The model size vary from 3 billion parameters up to 1.76 trillion parameters (e.g., GPT-4). The development of these models can be traced back to transformer-based pre-trained deep learning models like BERT (Devlin et al., 2018). Unlike the earlier version of deep learning models limited to text generation and classification, these models have shown increased performance in generating more context-aware text responses.

In-context instruction learning, one of the main features of LLMs, allows these models to apply knowledge from pre-trained data based on the context embedded within the input (the prompt), without requiring supervised training or fine-tuning with vast amounts of domain-specific data (e.g., over 100k text documents with annotated labels). This approach contrasts with traditional natural language processing methods where the limited size of the corpus has to be collected before the generation and task-specific labeled datasets like sentiment analysis or translation. LLMs have shifted toward more adaptive and comprehensive language generation tasks, such as assisting creative writing (e.g., Simon & Muise, 2022; Yuan et al., 2022; Zhao et al., 2023).

Due to the high quality of generated output with models, it is easy to make people believe that these models can 'understand' users' intentions. This perceived understanding has gathered attention from various fields, including sociology, medicine, and psychology. Recent studies have begun to explore the cognitive abilities of these artificial agents (Bubeck et al., 2023; Momennejad et al., 2023). This exploration includes understanding (West et al., 2023), reasoning (Li et al., 2023; Qiu et al., 2023), emotion recognition (Yongsatianchot et al., 2023), bias (Gallegos et al., 2023), and general intelligence (Bubeck et al., 2023; Momennejad et al., 2023).

Eliciting Reasoning in Large Language Models

While LLMs have shown impressive performance in generating human-like text, these generative models still suffer from many technical issues like a hallucination that deteriorates their accurate understanding of user intention and trust. Furthermore, functional limitations hinder their applicability in domains requiring automated social and emotional support. For example, instructing these LLMs to generate empathic responses, the outputs from these models often take the form of 'advising' or 'listing facts and action items,' lacking attention to detail and specificity, leading to the impression that the model does not truly understand what the user is conveying and perceived less empathic.

Recent developments in prompt engineering by eliciting reasoning steps in solving natural language tasks like arithmetic reasoning have shown effective in improving performance (Kojima et al., 2022; Prystawski et al., 2022; Wei et al., 2022; Yao et al., 2023). CoT prompting, a method that encourages the model to 'reason' through a problem before generating a response, has significantly improved problem-solving tasks (Kojima et al., 2022) and metaphor comprehension (Prystawski et al., 2022), paving the new direction for generative models' real-world application like medical diagnosis (Singhal et al., 2023). However, the effectiveness of eliciting reasoning steps in generating socially appropriate responses for improved social support remains under-explored. Consequently, there is a need for distinct approaches in LLMs to reason about emotions and situations.

The Present Study

We investigated whether eliciting empathic reasoning in LLMs leads to natural responses. Therefore, we developed a CoE prompting to reason emotion and situational factors that could help the model to accurately infer the client's emotional experience in mental healthcare and thus choose the most appropriate and context-aware empathy strategy to communicate.

Methods

Prompting Methods

We used the GPT-3.5-Turbo model (identified as 'text-DaVinci-003') released by OpenAI[1], which was the most recent and outperformed previous generative models up to a point. To generate text, four main hyperparameters were configured*: temperature, top p parameter, frequency penalty,* and *presence penalty. Temperature* controls the randomness of the output; a higher temperature results in more varied text, whereas a lower temperature produces more monotonous output. We set it to 0.9 to foster the creation of diverse text that aligns with the different purposes and contents of each CoE prompt. The *top p* parameter, also known as nucleus sampling, controls the cumulative probability threshold at which the model's predictions are considered; we set it to 1, meaning all tokens are considered regardless of their probability. This approach was chosen to generate the most likely next-word

[1] https://openai.com

predictions. The *frequency penalty* reduces the model's likelihood of repeating the same lines or phrases, enhancing the uniqueness of the generated text; the *presence penalty* discourages the repetition of previously mentioned tokens, leading to a broader exploration of topics. We set the frequency penalty to 0, indicating no penalty for repetition, and the presence penalty to 0.6 to encourage topic variety without overly penalizing relevant repetitions. These settings were chosen to balance creativity with coherence, and considering conventionally used settings in previous findings, ensuring the generated text was diverse yet contextually relevant. We used the open-source Python package, LangChain library, for data generation and automatic evaluation. The library optimizes model inference at scale by including code templates for prompting, instruction-tuning, and chaining prompts.

Chain-of-Empathy Prompting

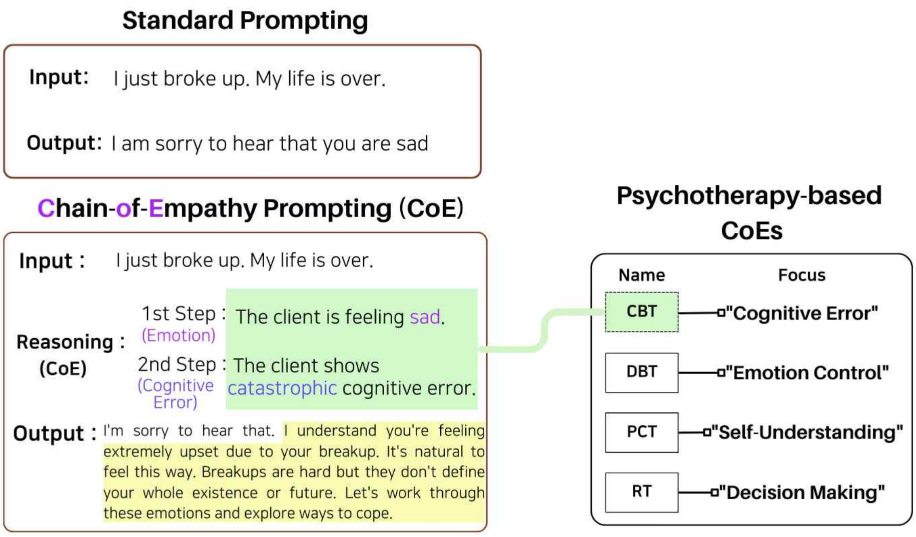
Figure 1 and Table 1 show four unique prompts that reflect the reasoning steps needed for empathic response generation in addition to the baseline prompting condition (no reasoning step): Cognitive-Behavioral Therapy (CBT; Hofmann et al., 2010; Kaczkurkin & Foa, 2022), Dialectical Behavior Therapy (DBT; Linehan, 1987), Person-Centered Therapy (PCT; Cooper & McLeod, 2011; Knutson & Koch, 2022), and Reality Therapy (RT; Arab & Khodabakhshi-Koolaee, 2022; Wubbolding et al., 2017). Except for the base condition, these prompts' instructions were designed to reflect the therapists' reasoning style in their counseling methods.

For the CBT-CoE prompt, the instructions were, "Reflect on how the client's negative thought patterns may be contributing to their emotions. Contemplate alternative perspectives to their interpretations of the situation that might alleviate emotional distress", along with definitions of potential negative thought patterns identified by CBT. These include: Dichotomous thinking ('viewing situations in extreme all-or-nothing terms'), catastrophizing ('anticipating the worst possible outcome'), overgeneralization ('drawing broad conclusion based on a single incident or piece of evidence'), mindreading ('believing one can predict other persons' thoughts or feelings without concrete evidence or communication') and self-blaming ('ascribing every negative event or result to oneself without considering other factors'). For the DBT-CoE prompt, the instruction was, "Identify the emotions the client is struggling to regulate from their message. Understand how these emotions reflect the client's current ability to handle distress or intense feelings." The instruction for PCT-CoE prompt was, "Identify the client's emotional state from their message. The client's words could indicate their current self-understanding or self-perception."; The instruction for RT-CoE prompt was, "Identify any expressed unmet needs or dissatisfactions in the client's message. Understand and reflect on how these needs or dissatisfactions could relate to the client's current situation or their struggle to fulfill basic needs for love and belonging. Avoid focusing on diagnosing mental health issues or problems; instead, aim to help the client take control of their actions and make wise decisions." Due to sensitivity to specific phrases and words written in the prompt instruction of LLMs, we made sure the outputs convey the same context and goal by paraphrasing each prompt moderately multiple times to achieve consistent output.

Models in each prompting condition were tested in zero-shot learning initially, with only instructions on which option to choose per class: empathy strategy ('emotional reaction', 'exploration', and 'interpretation') and communication level ('no expression', 'weak', and 'strong') that were originally labeled from the benchmark data (Sharma et al., 2020). For our study, we designed prompts for LLMs to respond based on these reasoning steps involved in each CoE condition: (1) identify any word that represents the client's emotion, and (2) understand the individual/situational factors that may have led to the expression in the client's message.

<Table 1> Comparison of goals and reasoning style in different psychotherapy based CoEs.

	Prompt Conditions			
	CBT-CoE	DBT-CoE	PCT-CoE	RT-CoE
Goals	Tackling negative thought patterns	Addressing emotional dysregulation	Encouraging self-understanding	Identifying cause of the dissatisfaction with reality and help tackle goals
Reasoning steps	1) Find negative thought patterns if present in users' messages 2) Reason why that negative thought derived from	1) Identify the client's emotional state 2) Interpret the client's words to understand their struggles with emotional regulation	1) Identify the client's emotional state from their message 2) Interpret the client's words to understand their current level of self-understanding or self-perception	1) Convey empathic understanding, unconditional positive regard, and congruence 2) Understand how these needs or dissatisfactions reflect the client's current situation or inability to meet their basic needs for love and belonging


Standard Prompting
Input: I just broke up. My life is over.
Output: I am sorry to hear that you are sad
Chain-of-Empathy Prompting (CoE)
Input : I just broke up. My life is over.
Reasoning : (CoE)
1st Step : (Emotion) The client is feeling sad.
2nd Step : (Cognitive Error) The client shows catastrophic cognitive error.
Output : I'm sorry to hear that. I understand you're feeling extremely upset due to your breakup. It's natural to feel this way. Breakups are hard but they don't define your whole existence or future. Let's work through these emotions and explore ways to cope.
Psychotherapy-based CoEs
Name
Focus
CBT "Cognitive Error"
DBT "Emotion Control"
PCT "Self-Understanding"
RT "Decision Making"

(Figure 1) CoE prompting with cognitive reasoning of human's emotion based on psychotherapy models

Experiments

We instructed the LLM to generate appropriate responses to posts from Reddit seeking advice on personal issues causing stress and to predict the most suitable empathic expression strategy. For the ground truth label of each empathy strategy class, we utilized the EPITOME dataset, which comprises crowdsourced Reddit posts related to mental health, with an average inter-annotator agreement reported as above 0.68 (Sharma et al., 2020). This dataset and our computational approach to expressed empathy align with our aim to identify and characterize expressed empathy in texts.

The dataset included pairs of help-seeking and responding posts, each pair labeled based on (1) the expressed empathy (here, we reframed as 'empathy strategy' as its way of communicating empathy) and (2) the presence and 'level'of each expressed empathy (i.e., communication strength). The three empathy strategies—emotional reaction, exploration, and interpretation—correspond to levels 0, 1, and 2, and were identified from over 10,000 posts. While the original classifications are not exhaustive, they were adapted more comprehensively for an effective computational framework.

Following this adaptation and incorporating insights from recent psychology approaches (Weisberg et al., 2023), emotional reaction is defined as responses where the peer supporter expresses concern, compassion, or warmth towards the support seeker. Exploration involves delving into the seeker's unexpressed feelings, whereas interpretation entails communicating an understanding of the feelings and experiences inferred from the seeker's post. Notably, exploration and interpretation are associated with cognitive empathy, such as perspective-taking, while emotional reaction is linked to affective empathy, like empathic concern.

We applied the three empathy strategies within the framework of LLM classifiers designed as part of our proposed CoE prompts. This model does not directly equate to the empathy strategies themselves but serves as targets of classification. These particular empathy strategies were selected by the availability of a corpus annotated with empathy components, rendering it uniquely suitable for our study's objectives. Each psychotherapy model implemented in the CoE prompts can potentially utilize one or all of these empathy strategies, reflecting the multifaceted nature of empathy. This flexibility underscores the relevance of our approach to real-world psychotherapeutic settings, where varying expressions of empathy are crucial components of active listening and effective support.

In addition, pairs labeled level 0, indicating no expression of empathy, were excluded. The number of pairs for each strategy was as follows: emotion reaction=1,047, exploration=481, and interpretation=1,436. We randomly sampled 500 pairs in each emotional reaction and interpretation data to balance the number of pairs between strategies. Each strategy's final number of pairs was emotional reaction=500, exploration=480, and interpretation=500.

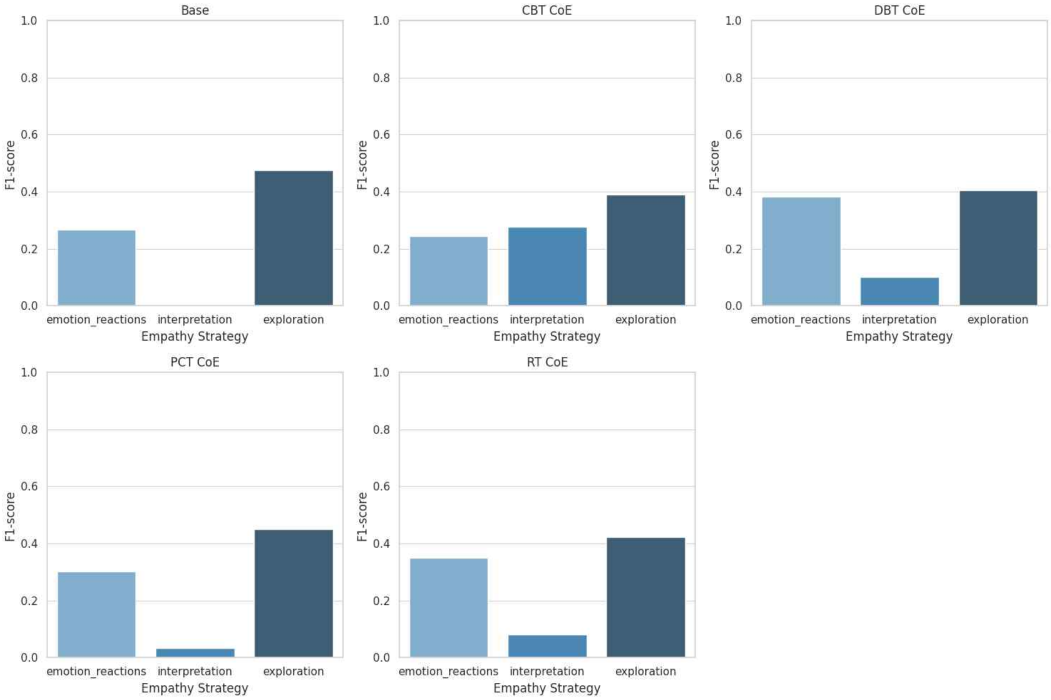
Results

Automatic Evaluation

Table 2 and Figure 2 show the performance of the empathy strategy classification of LLMs with each CoE prompt. Upon generating a response, each model with CoE prompts predicted which empathy strategy is most suitable for each seeker's post among the three strategies. We then compared the predicted empathy strategy with the ground truth empathy strategy and calculated prediction accuracy. We employed multiple metrics to evaluate the effectiveness of the model in categorizing empathy strategies across different prompts. The statistical metrics we used - accuracy, precision, recall, and F1-score - are commonly used in fields that employ computational methods to assess model performance such as information retrieval and natural language processing (Powers, 2011; Tharwat, 2020). Each metric offers valuable insights into the model's performance.

<Table 2> Model performance in empathy strategy classification task by CoE prompting conditions.

	Acc.	Emotional Reaction			Interpretation			Exploration		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Base	0.340	0.467	0.185	0.27	0	0	0	0.327	0.866	0.475
CBT-CoE	0.319	0.463	0.165	0.244	0.293	0.260	0.276	0.303	0.543	0.389
DBT-CoE	0.334	0.392	0.372	0.382	0.291	0.060	0.100	0.309	0.582	0.404
PCT-CoE	0.336	0.399	0.243	0.302	0.333	0.016	0.031	0.319	0.757	0.449
RT-CoE	0.336	0.407	0.308	0.350	0.354	0.044	0.79	0.309	0.664	0.420



(Figure 2) Empathy strategy classification accuracy per prompt conditions. Compared to Base condition, CBT-CoE provided the balanced set of each empathy expression but less emotional reaction than other CoEs.

Accuracy measures the proportion of total predictions that were correct. Within the context of our experiment, accuracy serves as a metric to assess the level of agreement between the model predictions and actual empathy strategies employed. However, it should be noted that accuracy alone may not be sufficient to capture the complexity of our experimental design. *Precision* assesses the correctness of positive predictions in relation to total positive predictions made. *Recall* evaluates the ability of the model to identify all relevant instances of a specific empathy strategy. A high recall indicates that the model effectively captures instances of a particular strategy. It is crucial in our study to identify which empathy strategies are better recognized under different prompting conditions. *F1-score* provides a mean of precision and recall, balancing the two metrics, which helps us understand how well each model balances these aspects in predicting empathy strategies. This is important for our study, where class distributions are imbalanced, and both recall and precision of empathy strategies are crucial.

The evaluation results reveal that overall accuracy for each strategy under different prompt conditions is relatively low, which was anticipated given our study's exploratory nature. Unlike traditional machine learning approaches, our research focuses on empathy strategy classification under varied prompts. The goal is to understand how different prompts influence the model's preference rather than just exceeding an accuracy benchmark. Therefore instead of focusing on achieving high absolute scores, we compared relative accuracy (i.e., which empathic accuracy is relatively higher in which prompt conditions). In specific, Emotional Reaction was most effectively captured by the DBT-CoE model, achieving the highest F1 score (0.382) among the conditions. This suggests a nuanced understanding and classification of emotional reaction within the DBT-CoE framework. Interpretation prediction was highest with the CBT-CoE model (F1 score=0.276) indicating a relatively better alignment with CBT's cognitive process-focused approach. Notably, all CoE prompts outperformed the baseline in predicting interpretation, demonstrating the added value of CoE prompting in enhancing model sensitivity to interpretative empathy. Exploration was best identified by the baseline condition, with an F1 score of 0.475, suggesting a general tendency of the model without specific CoE prompting to favor exploration strategies. However, among the CoE conditions, PCT-CoE came closest to this benchmark (F1 score=0.449), indicating its relative strength in eliciting exploration-oriented empathic responses.

Outputs with errors in the predicted strategy names were excluded from the analysis. Most of these errors resulted from the nature of LLM as a generative model, which behaves differently from traditional supervised learning models for classification tasks. Despite explicit instructions in the prompt, the models occasionally generated 'noise' output and predicted strategies that were not among the provided options. These errors include responses of failed predictions or response retrieval (e.g., "No Empathy Strategy"). In addition, they sometimes predicted new strategies that did not fall into any of the predefined three strategies (e.g., "Reflection,""Validation: acknowledging the client's feelings and experiences," and "Approval: expressing approval or positive reinforcement to the client").

Qualitative Evaluations of AI Alignment

We conducted qualitative analysis to investigate whether the results produced by LLMs precisely follow the instructions provided in the prompts and align with the intentions as perceived and evaluated by human judges. Our aim was not to perform an exhaustive analysis, but rather to gauge the quality of the outputs in a qualitative and descriptive manner to help interpret model performance. In general, LLMs often fail to address the main objective and instruction intended by the researcher (or user), despite generating 'looking good' responses at first glance.

When compared to the benchmark dataset (Sharma et al., 2020), while many human peer supporters often provided brief comments and shared personal opinions or gave advice, the CoE LLMs mostly responded with at least two empathic strategies and frequently ended their response with suggestions for seeking professional help. In addition, we found a pattern of LLMs' responses that the model usually initiated by interpreting users' current state and proceed to provide advice or exploring potential options. For example, when a distressed seeker could not control her anxiety after a violent fight between her parents, DBT-CoE prompt responded with multiple empathic strategies, "I am so sorry you had to witness that. Understandably, you're feeling overwhelmed and scared right now. It's not okay for anyone to threaten or hurt another person, and it's not your fault. How can I support you right now?" This contradicts the original human response in benchmark data: "Everything is wrong with people."

Our specific criteria were whether LLMs generate responses that align with 1) prompt instruction 2) the goal of each CoE prompt by using manual evaluation (Table 3 and 4). Two evaluators with a psychology background (who are educated in undergraduate-level psychology with background knowledge of each psychotherapy model and large language models). The selection of evaluators and implementation of a structured evaluation process was designed to obtain qualitative insights into the LLM's performance, particularly its functionality to align generated responses with the instructions provided in different CoE prompt conditions. Our graduate-level psychology authors conducted training sessions that improved evaluators' knowledge of psychology. The training was specific to an in-depth investigation of LLM's alignment accuracy. This approach aimed not to classify responses into specific

<Table 3> Total percentage (counts) of valid responses remaining after each AI alignment evaluation step

Prompt Condition	Empathy Strategy	First Step- Intent Error Filter	Second Step- Instruction-following	Third Step- Reasoning
CBT-CoE	Emotional Reaction	41.21 (204/495)	84.00 (84/100)	84.52 (71/84)
	Exploration	42.00 (197/469)	95.00 (95/100)	89.50 (85/95)
	Interpretation	43.74 (213/487)	86.00 (86/100)	83.90 (72/86)
DBT-CoE	Emotional Reaction	74.95 (368/491)	97.00 (97/100)	94.85 (92/97)
	Exploration	74.68 (354/474)	98.00 (98/100)	97.96 (96/98)
	Interpretation	72.32 (358/495)	96.00 (96/100)	94.79 (90/96)
PCT-CoE	Emotional Reaction	56.15 (274/488)	95.00 (95/100)	96.84 (92/95)
	Exploration	57.96 (273/471)	96.00 (96/100)	95.83 (92/96)
	Interpretation	56.88 (273/480)	99.00 (99/100)	94.95 (94/99)
RT-CoE	Emotional Reaction	33.06 (162/490)	94.00 (94/100)	96.80 (91/94)
	Exploration	33.54 (160/477)	92.00 (92/100)	98.90 (91/92)
	Interpretation	31.49 (154/489)	87.00 (87/100)	100 (87/87)

clinical symptoms or populations but to identify areas where the system may not fully follow the researchers' prompts, such as hallucination, generating false information, or simple errors. Thus, we considered this level of evaluator expertise suitable for the initial depth of assessment required in this exploratory phase. However, for future work, it is recommended that formal inter-rater reliability measures be included to enhance the robustness of human evaluation. Including evaluations with professional therapists would greatly enhance its applicability in related fields. The evaluation was conducted in three-fold steps:

(1) First Step: The first step was to automatically count simple errors that were detectable by defining filter keywords ("{emotion}" instead of actual emotion inferred). For CBT-CoE, the model identified cognitive errors based on the client's relevant statements. Similarly, for DBT-CoE, PCT-CoE, and RT-CoE, the focus was on inferring emotions and psychological experiences from the client's words. Data lacking clear outcomes or failing to reflect this reasoning about the client's emotional experience were excluded. Valid data included clear reasoning about the client's psychological state.

(2) Second Step: The aim of the second step evaluation was to evaluate the LLM's alignment with the prompt instructions, specifically whether it generated responses that can be characterized as

empathic. To do this, we randomly sampled 100 data points for each CoE prompt and corresponding empathy strategy. Human evaluators then assessed whether the LLM generated responses processed the instructed reasoning process for each psychotherapy approach. Two CoE prompt responses were assigned per person (CBT and RT; DBT and PCT) and assessed whether the response generated by the LLM adhere to the instructed reasoning process, regardless of the content of the final response. Responses that aligned with the instructions were annotated '1,' while those that did not were annotated '0.'

(3) Third Step: The final step involved evaluating the LLM's performance in aligning with the objectives of each CoE prompt, specifically examining how well the LLM's responses corresponded to the intended goals of each psychotherapy model. We focused on the model's ability to accurately process and reflect the client's emotional state and situation in its generated responses. Each data was annotated with '1' for correct alignment with prompt's objectives, and '0' for those that fail to align with it (with explanations also provided by the human evaluators).

Based on the third-step evaluation results, additional behavioral coding was performed to understand why human evaluators assigned a '0' rating to specific data. Generally, the LLM's final responses

<Table 4> Evaluation of human alignment steps of CoE results (reasoning steps and responses)

Human Alignment Evaluation Steps		
Steps	Goals	Examples of Misalignment Cases
First step	Filter out technical error (0: incorrect, 1: correct)	Reasoning: "the client is feeling {emotion}."
Second step	Evaluate the prompt instruction following (0: does not generate reasoning based on the psychotherapy goal, 1: generate reasoning based on the psychotherapy goal)	Support seeker: "My head is exploding.. All of this. How do you cope with this life?" CBT-CoE response: "The client is feeling overwhelmed."(not a negative thought pattern)
Third step	Evaluate quality of the response according to the each psychotherapy methods' goal and reasoning style	Support seeker: "I miss my mum. I'm just crying a lot today and. I'm really depressed because I miss my mum." reasoning: "the client is feeling sadness the client says 'I miss my mum' which could reflect the cognitive error of overgeneralization that influence negative thought patterns." (not a case of 'overgeneralization')

were generic and inclusive, which did not seem to include any problematic content. However, the misalignment error usually found at the two reasoning steps. The first reason for a '0' rating was the LLM's incorrect inference of the client's emotional states and thoughts. The second reason was when the LLM did not understand the client's posts. For example, when the client used self-mocking utterances, the bot could not grasp the underlying meaning of the post. The LLM sometimes provided the final response, including something irrelevant to the client's post. The rationale for classification in the third step was to find cases when the client's post did not include any negative content (or negative thought pattern). However, the LLM regarded the client's statements as negative or unfortunate.

The rationale for incorrect labeling varied across different psychotherapy approaches. In the CBT-CoE condition, the system often inaccurately interpreted the client's negative thoughts. The DBT-CoE frequently misunderstood the client's emotions and thoughts, including the underlying causes of emotional dysregulation. The PCT-CoE often misinterpreted the client's situations by fixating on particular words in the posts, such as "trust," "wish," and "funeral." Despite the fact that the posts did not explicitly address the client's need for a sense of belonging, the RT-CoE effectively addressed client's messages using the RT counseling method.

Discussion

In this study, we hypothesize that eliciting empathic reasoning steps from large language models based on distinct psychotherapy models would demonstrate varying preferences and accuracies in generating empathic responses, thereby exploring the empathic response generation ability of generative AIs. Our findings support this hypothesis, revealing that LLMs without specific reasoning prompts significantly prefer the exploration strategy, with interpretation being the least favored.

Despite all reasoning prompts predominantly generating exploration-related responses, they diverged from the base (non-reasoning) prompt by generating the interpretation strategy to varying degrees. Only the CBT-CoE prompt led to the highest frequency of interpretation strategy responses. This outcome likely mirrors the fundamental objective of Cognitive-Behavioral Therapy (CBT), which focuses on identifying and elucidating cognitive distortions clients express. These results underscore the critical role of context-specific empathic reasoning when integrating generative AI within therapeutic interactions. This variation in classification accuracy across different CoE conditions underscores the importance of tailored empathic reasoning in enhancing the relevance and impact of AI-generated empathic communications.

We suggest several remaining points that were not addressed because they were outside our scope but should be considered for future research. First, given empathy's multifaceted property and subjectivity, a wider range of evaluative criteria for AI alignment in empathic response will provide us with more insight into LLM's ability to express empathy that aligns with personal interest and situational context. Our model mainly targets the general population from online peer support due to the limited availability of benchmark datasets for testing our models. As a result, we did not consider incorporating a broader spectrum of client data, such as symptom types and severity levels (e.g., normal, neurotic, borderline, psychotic). Future studies should incorporate expert-based evaluation to handle data from varied mental health conditions for effective AI-based clinical support (Ayers et al., 2023). At the same time, safety and confidentiality concerns should also be addressed when handling the data (Gottlieb & Silvis, 2023).

Our evaluation focused solely on the empathic accuracy of the LLMs' and its alignment with prompt instruction, but did not measure user perception of generated output. User perception of empathic expression varies depending on whether they interact with humans or artificially intelligent systems (Medeiros et al., 2021). Furthermore, people perceive and react differently to AIs' empathic expressions (Urakami et al., 2019). Thus, future works should investigate how users perceive and respond to the models' empathic responses to enhance our understanding of the efficacy of LLMs' empathic expressions.

For human evaluation of AI alignment, we used a single LLM model (GPT-3.5-Turbo) and one domain, mental health. Incorporating a diverse text corpus, including career coaching and motivational interviewing (Miller & Rollnick, 2012), could enable LLMs to produce more personalized communication. This presents an opportunity for future research to encompass a wider array of topics and conversational styles, thereby increasing the reliability of LLM's performance. Additionally, different LLMs may excel in varied capabilities, leading each LLM to display optimal performance in specific tasks (Sivarajkumar et al., 2023). Investigating and assessing the empathic expressions generated by different LLMs is crucial for a comprehensive evaluation of LLMs' ability to discern human emotions and craft appropriate, empathic responses.

References

Ahn, Y., Zhang, Y., Park, Y., & Lee, J. (2020). A chatbot solution to chat app problems: Envisioning a chatbot counseling system for teenage victims of online sexual exploitation. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-7.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). American Psychiatric Publishing, Inc.

Anderson, C., & Keltner, D. (2002). The role of empathy in the formation and maintenance of social bonds. *Behavioral and Brain Sciences*, *25*(1), 21-22.

Arab, A., & Khodabakhshi-Koolaee, A. (2022). The magic of WDEP in reality therapy: Improving intimacy needs and personal communication in married males. *European Journal of Psychology Open*, *81*(3), 97 – 103.

Asada, M. (2015). Towards artificial empathy: how can artificial empathy follow the developmental pathway of natural empathy?. *International Journal of Social Robotics*, *7*, 19-33.

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. Oxford, England: International University Press.

Brocki, L., Dyer, G. C., Gladka, A., & Chung, N. C. (2023). Deep learning mental health dialogue system. *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 395-398.

Bommarito II, M., & Katz, D. M. (2022). GPT takes the bar exam. *arXiv preprint arXiv:2212.14402*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Buechel, S., Buffone, A., Slaff, B., Ungar, L., & Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4758-4765.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023).

Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv: 2303.12712.*

Casas, J., Spring, T., Daher, K., Mugellini, E., Khaled, O. A., & Cudré-Mauroux, P. (2021). Enhancing conversational agents with empathic abilities. *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, Online,* 41-47.

Charrier, L., Rieger, A., Galdeano, A., Cordier, A., Lefort, M., & Hassas, S. (2019). The rope scale: a measure of how empathic a robot is perceived. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 656-657.

Concannon, S., & Tomalin, M. (2023). Measuring perceived empathy in dialogue systems. *AI & SOCIETY*, 1-15.

Cooper, M., & McLeod, J. (2011). Person-centered therapy: A pluralistic perspective. *Person-Centered & Experiential Psychotherapies*, *10*(3), 210-223.

Davis, M. A. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalogue of Selected Documents in Psychology*, *10*, 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113 – 126.

De Vignemont, F., & Singer, T. (2006). The empathic brain: how, when and why?. *Trends in Cognitive Sciences*, *10*(10), 435-441.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders*, *6*(1), 249-262.

Eisenberg, N. (2014). *Altruistic emotion, cognition, and behavior (PLE: Emotion).* Psychology Press.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), e19.

Gabriel, I. (2020). Artificial Intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411-437.

Glasser, W. (1965). *Reality therapy.* New York: Harper & Row.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2023). Bias and fairness in large language models: A survey. *arxiv preprint arXiv:2309.00770.*

Gottlieb, S., & Silvis, L. (2023). How to safely integrate large language models Into health care. *JAMA Health Forum, 4*(9), e233909.

Gruschka, F., Lahnala, A., Welch, C., & Flek, L. (2023). Domain transfer for empathy, distress, and personality prediction. *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Canada,* 553-557.

Hall, J. A., & Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology*, *159*(3), 225-243.

Herrera, F., Bailenson, J., Weisz, E., Ogle, E., & Zaki, J. (2018). Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS one*, *13*(10), e0204494.

Hofmann, S. G., Sawyer, A. T., & Fang, A. (2010). The empirical status of the "new wave" of cognitive behavioral therapy. *Psychiatric Clinics*, *33*(3), 701-710.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, *6*(11), e12106.

Kaczkurkin, A. N., & Foa, E. B. (2022). Cognitive-behavioral therapy for anxiety disorders: an update on the empirical evidence. *Dialogues in Clinical Neuroscience*, *17*(3) 337-346.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Knutson, D., & Koch, J. M. (2022). Person-centered therapy as applied to work with transgender and gender diverse clients. *Journal of Humanistic Psychology*, *62*(1), 104-122.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lee, B., & Yi, M. Y. (2023). Understanding the empathetic reactivity of conversational agents: Measure development and validation. *International Journal of Human–Computer Interaction*, 1-19.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J. G., & Chen, W. (2023). Making language models better reasoners with step-aware verifier. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, *1*, 5315-5333.

Linehan, M. M. (1987). Dialectical behavioral therapy: A cognitive behavioral approach to parasuicide. *Journal of Personality Disorders*, *1*(4), 328-333.

Linehan, M. M., Armstrong, H. E., Suarez, A., Allmon, D., & Heard, H. L. (1991). Cognitive-behavioral treatment of chronically parasuicidal borderline patients. *Archives of General Psychiatry*, *48*(12), 1060-1064.

Medeiros, L., Bosse, T., & Gerritsen, C. (2021). Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. *IEEE Transactions on Human-Machine Systems, 52*(3), 343-353.

Mehta, A., Niles, A. N., Vargas, J. H., Marafon, T., Couto, D. D., & Gross, J. J. (2021). Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. *Journal of Medical Internet Research*, *23*(6), e26771.

Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change and grow*. Guilford press.

Momennejad, I., Hasanbeig, H., Vieira, F., Sharma, H., Ness, R. O., Jojic, N., Palangi, H., & Larson, J. (2023). Evaluating cognitive maps and planning in large language models with CogEval. *arXiv preprint arXiv:2309.15129*.

Neff, K. (2003). Self-compassion: An alternative conceptualization of a healthy attitude toward oneself. *Self and Identity*, *2*(2), 85-101.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Nwosu, A., Boardman, S., Husain, M. M., & Doraiswamy, P. M. (2022). Digital therapeutics for mental health: Is attrition the Achilles heel?. *Frontiers in Psychiatry*, *13*, 900615.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 1-40.

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37-63.

Prystawski, B., Thibodeau, P., Potts, C., & Goodman, N. D. (2022). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.

Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., ... & Ren, X. (2023). Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement. *arXiv preprint arXiv:2310.08559*.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2018). Towards empathic Open-domain Conversation Models: a New Benchmark and Dataset. *arXiv preprint arXiv:1811.00207*.

Rasouli, S., Gupta, G., Nilsen, E., & Dautenhahn, K. (2022). Potential applications of social robots in robot-assisted interventions for social anxiety. *International Journal of Social Robotics*, *14*(5), 1-32.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, *21*(2), 95.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural Psychology*, *32*(1), 76-92.

Sharma, A., Miner, A., Atkins, D., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online,* 5263–5276.

Simon, N., & Muise, C. (2022). TattleTale: Storytelling with Planning and Large Language Models. *ICAPS Workshop on Scheduling and Planning Applications.*

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180.

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2023). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. *arXiv preprint arXiv:2309.08008.*

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., ... & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of Medical Internet Research*, *22*(3), e16235.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models.* https://crfm.stanford.edu/2023/03/13/alpaca.html.

Tharwat, A. (2020). Classification assessment methods. *Applied computing and informatics*, *17*(1), 168-192.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.*

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.*

Truax, C. B., & Carkhuff, R. (2007). *Toward effective counseling and psychotherapy: Training and practice.* Transaction Publishers.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59,* 433–460.

Urakami, J., Moore, B. A., Sutthithatip, S., & Park, S. (2019). Users' perception of empathic expressions by an advanced intelligent system. *Proceedings of the 7th International Conference on Human-Agent Interaction, Japan*, 11-18.

Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Yu, M., Ma, X., & Wang, H. (2021). Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1-31.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903.*

Weisberg, O., Daniels, S., & Bar-Kalifa, E. (2023). Emotional expression and empathy in an online peer support platform. *Journal of Counseling Psychology*, *70*(6), 671.

Welivita, A., Xie, Y., & Pu, P. (2021). A large-scale dataset for empathic response generation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online and Dominican Republic*, 1251-1264.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., ... & Choi, Y. (2023). The generative AI paradox:"What it can create, it may not understand". *arXiv preprint arXiv:2311.00059*.

Wondra, J. D., & Ellsworth, P. C. (2015). An appraisal theory of empathy and other vicarious emotional experiences. *Psychological Review*, *122*(3), 411-428.

Wubbolding, R. E., Casstevens, W. J., & Fulkerson, M. H. (2017). Using the WDEP system of Reality Therapy to support person centered treatment planning. *Journal of Counseling & Development, 95*(4), 472-477.

Xie, B., & Park, C. H. (2021). Empathic robot with transformer-based dialogue agent. *2021 18th International Conference on Ubiquitous Robots (UR), Korea,* 290–295.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601.*

Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: story writing with large language models. *27th International Conference on Intelligent User Interfaces,* Finland, 841-852.

Yongsatianchot, N., Torshizi, P. G., & Marsella, S. (2023). Investigating large language models' perception of emotion using appraisal theory. *arXiv preprint arXiv:2310.04450.*

Zaki, J. (2019). *The war for kindness: Building empathy in a fractured world.* Crown.

Zhang, S. J., Florin, S., Lee, A. N., Niknafs, E., Marginean, A., Wang, A., ... & Drori, I. (2023). Exploring the MIT mathematics and EECS curriculum using large language models. *arXiv preprint arXiv:2306.08997.*

Zhao, Z., Song, S., Duah, B., Macbeth, J., Carter, S., Van, M. P., ... & Filipowicz, A. L. (2023). More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. *Proceedings of the 15th Conference on Creativity and Cognition, Online*, 368-370.

Zhou, L., Gao, J., Li, D., & Shum, H. Y. (2020). The design and implementation of xiaoice, an empathic social chatbot. *Computational Linguistics*, *46*(1), 53-93.

1차 원고 접수: 2023. 12. 15
1차 심사 완료: 2024. 02. 05
2차 원고 접수: 2024. 03. 04
2차 심사 완료: 2024. 03. 14
최종 게재 확정: 2024. 03. 20

(요 약)

인공지능 기반 사회적 지지를 위한 대형언어모형의 공감적 추론 향상: 심리치료 모형을 중심으로

이윤경[1)] 이인주[1)] 신민정[2)] 배서연[1)] 한소원[1,2)]

[1)]서울대학교 심리학과 [2)]서울대학교 협동과정 인지과학전공

대형언어모형(LLM)을 현실에 적용하려는 지속적인 노력에도 불구하고, 인공지능이 맥락을 이해하고 사람의 의도에 맞게 사회적 지지를 제공하는 능력은 아직 제한적이다. 본 연구에서는 LLM이 사람의 감정 상태를 추론하도록 유도하기 위해, 심리 치료 이론을 기반으로 한 공감 체인(Chain of Empathy, CoE) 프롬프트 방법을 새로 개발했다. CoE 기반 LLM은 인지-행동 치료(CBT), 변증법적 행동 치료(DBT), 인간 중심 치료(PCT) 및 현실 치료(RT)와 같은 다양한 심리 치료 방식을 참고하였으며, 각 방식의 목적에 맞게 내담자의 정신 상태를 해석하도록 설계했다. CoE 기반 추론을 유도하지 않은 조건에서는 LLM이 사회적 지지를 구하는 내담자의 글에 주로 탐색적 공감 표현(예: 개방형 질문)만을 생성했으며, 추론을 유도한 조건에서는 각 심리 치료 모형을 대표하는 정신 상태 추론 방법과 일치하는 다양한 공감 표현을 생성했다. 공감 표현 분류 과제에서 CBT 기반 CoE는 감정적 반응, 탐색, 해석 등을 가장 균형적으로 분류하였으나, DBT 및 PCT 기반 CoE는 감정적 반응 공감 표현을 더 잘 분류하였다. 추가로, 각 프롬프트 조건 별로 생성된 텍스트 데이터를 정성적으로 분석하고 정렬 정확도를 평가하였다. 본 연구의 결과는 감정 및 맥락 이해가 인간-인공지능 의사소통에 미치는 영향에 대한 함의를 제공한다. 특히 인공지능이 안전하고 공감적으로 인간과 소통하는 데 있어 추론 방식이 중요하다는 근거를 제공하며, 이러한 추론 능력을 높이는 데 심리학의 이론이 인공지능의 발전과 활용에 기여할 수 있음을 시사한다.

주제어 : 공감, 인공지능 기반 사회적 지지, 대형언어모형, 자연어처리, 인지적 AI, AI Alignment