Let's Put {Speaker} into {Interlocutor}'s Shoes! Exploring the Impact of Zero-Shot Chain-of-Thought on Empathy

Anonymous ACL submission

Abstract

This study investigates the effectiveness of the Zero-shot Chain-of-Thought (CoT) approach, specifically the "*Let's think step by step.*", in enhancing both the perceived empathy and empathetic reasoning abilities of 14 Large Language Models (LLMs) in total. However, our experiments indicate that Zero-shot CoT does not significantly improve LLMs' empathetic reasoning compared to Zero-shot In-Context Learning (ICL), based on a range of performance metrics. Notably, we uncover that employing the perspective-taking prompt (i.e., "*Let's put speaker into interlocutor's shoes.*") strengthens empathetic reasoning, leading to notable improvements in perceived empathy scores. The source code will be made available upon publication.

1 Introduction

Recent studies have witnessed the success of Chain-of-Thought (CoT) prompting (Wei et al., 2022; Shi et al., 2022; Zhang et al., 2022) in achieving remarkable zero-/few-shot performance on complex reasoning tasks, including arithmetic, symbolic, and multi-modal (Zhang et al., 2023), which benefited from providing step-by-step reasoning, called *rationale*, into Large Language Models (LLMs) (Ouyang et al., 2022; Chowdhery et al., 2022; Chung et al., 2022; OpenAI, 2023a,b). In a zero-shot setting, a standard approach of Zero-shot CoT (Kojima et al., 2022) has demonstrated significant performance improvements simply using "*Let's think step by step.*" Previous work has explored the efficacy of Zero-shot CoT in enhancing zero-shot generalization performance on social knowledge tasks, such as social bias, toxicity (Shaikh et al., 2022), and Theory-of-Mind (ToM) (Moghaddam and Honey, 2023). Motivated by the prior work, we focus on exploring whether zero-shot CoT unlocks the empathetic reasoning capability of LLMs in terms of the social dialogue domain.

As highlighted in (Sap et al., 2022), for an AI assistant to be social and interactive, it should possess social reasoning capabilities, including empathy and understanding of the interlocutor's perspective. Empathy involves comprehending another individual's experiences, feelings, and thoughts in interpersonal communication. As effective listening in the communication process is significant (Main, 1985; Castleberry and Shepherd, 1993), it plays a crucial role in empathetic communication, referred to as Active Empathetic Listening (AEL) (Comer and Drollinger, 1999). The conceptual framework of AEL comprises three main dimensions: *Sensing*, *Processing*, and *Responding*. Within the *Processing*, it is crucial to understand, evaluate, interpret, and remember the interlocutor's implications, which is facilitated by perceiving their messages. We argue that the *Processing* part is highly correlated to the *perspective-taking* (Davis, 1983; Ruby and Decety, 2004; Kim et al., 2021), which is the act of perceiving and understanding another person's situation by putting ourselves in the other's shoes. Given the importance of *perspective-taking* in empathy, we question whether Zero-shot CoT truly induces effective empathetic reasoning. While Zero-shot CoT excels at sequentially generating rationales to address specific problems (e.g., mathematics), we believe that for empathetic reasoning, a profound understanding or interpretation of a conversation is more valuable than merely discerning its superficial or literal meaning.

In this study, we first explore the potential of Zero-shot CoT to enhance the LLM's ability to express empathetic response by measuring two aspects: *perceived empathy* and *empathetic reasoning*. Through our experiments, we demonstrate that Zero-shot CoT is less effective in unlocking both the *perceived empathy* and *empathetic reasoning* abilities of LLMs compared to Zero-shot

In-Context Learning (ICL), as measured by various metrics. Furthermore, we find that the *perspective-taking* prompting method (i.e., "*Let's put speaker into interlocutor's shoes.*") increases the empathetic reasoning ability, resulting in improvement on the *perceived empathy* performance.

In summary, our main contributions are as follows:

- This study is the first to investigate the efficacy of Zero-shot CoT in empathetic dialogue generation. Additionally, we introduce the Zero-shot Perspective-Taking (Z-Pers) prompting method (i.e., "*Let's put speaker into interlocutor's shoes.*") based on the Active Empathetic Listening (AEL) framework.
- We carry out a detailed analysis, assessing perceived empathy and empathetic reasoning abilities across 14 recent LLMs, encompassing both open-sourced and proprietary models.
- Through extensive experiments, we demonstrate that Z-Pers, which enhances the empathetic reasoning capability of LLMs, delivers superior zero-shot performance in the empathetic dialogue generation task.

2 Related Work

Chain-of-Thought Prompting. Recently, Chain-of-Thought (CoT) prompting (Wei et al., 2022; Shi et al., 2022; Zhang et al., 2022) has improved the zero-/few-shot performance across a range of complex reasoning tasks, including arithmetic, commonsense, symbolic, and logical reasoning. A key aspect of CoT prompting is the use of *rationale*, representing step-by-step reasoning. Previous studies have introduced a straightforward prompting method, called Zero-shot-CoT (Kojima et al., 2022), which involves simply providing rationale-trigger sentence "*Let's think step by step.*" into the LLMs, resulting in substantial improvements in zero-shot performance on various reasoning tasks. Beyond these tasks, recent studies have attempted to apply Zero-shot-CoT prompting to social knowledge tasks that require social reasoning, such as toxicity, social bias (Shaikh et al., 2022), and Theory-of-Mind (ToM) (Moghaddam and Honey, 2023). We scrutinize the potential for enhancing the empathetic reasoning, which is one of the social reasoning, of LLMs through the use of rationale in this work.

Empathetic Dialogue Generation. With the release of the EMPATHETICDIALOGUES dataset (Rashkin et al., 2018), many studies have proposed social dialogue generative agents, specifically to express empathy in social dialogues, by leveraging a mixture of experts (Lin et al., 2019), mimicking the interlocutor's emotions (Majumder et al., 2020), commonsense knowledge (Sabour et al., 2022), causality (Wang et al., 2022), and the Rational Speech Acts (RSA) framework (Kim et al., 2021). Furthermore, a recent study (Lee et al., 2022) has demonstrated that GPT-3 (Brown et al., 2020), in a zero-/few-shot setting, achieved better performance than Blender 90M (Roller et al., 2020) on the EMPATHETICDIALOGUES dataset using proposed in-context example selection methods based on emotional situation information. This study is the first to explore the effectiveness of Zero-shot CoT in terms of empathetic reasoning capability, standing apart from previous work that utilized in-context learning for the empathetic dialogue generation.

3 Methodology

3.1 Task Formulation

The empathetic dialogue generation task is to generate an empathetic response y by understanding the interlocutor's emotional situation for a given dialogue context x, which is formulated as follows:

$$p(y|x, \mathcal{M}) = \prod_{t}^{|y|} p(y_t|\mathcal{M}, x, y_1, \ldots, y_{t-1}) \quad (1)$$

where $\mathcal{M} = \{R, C, E\}$ and $\mathcal{P}(S) = \{\mathcal{M} : \mathcal{M} \subseteq S\}$. R and E denote rationale and emotional situations, respectively. $C = \{(x_j, y_j)\}_1^k$ represents in-context examples and k denotes the number of in-context examples. For example, in the zero-shot setting, we do not provide any in-context examples ($C = \emptyset$) to the LLMs. Given the aim of this work – to investigate the effect of Zero-shot CoT on the empathetic reasoning ability of LLMs – we set $\mathcal{M} = \{R\}$ and $C = \emptyset$.

3.2 Zero-shot Chain of Thought

The Zero-shot Chain of Thought (Zero-shot CoT) consists of two-stage prompting: (1) Reasoning Extraction and (2) Answer Extraction. Each stage of Zero-shot CoT is briefly described below, including its application for empathetic reasoning.

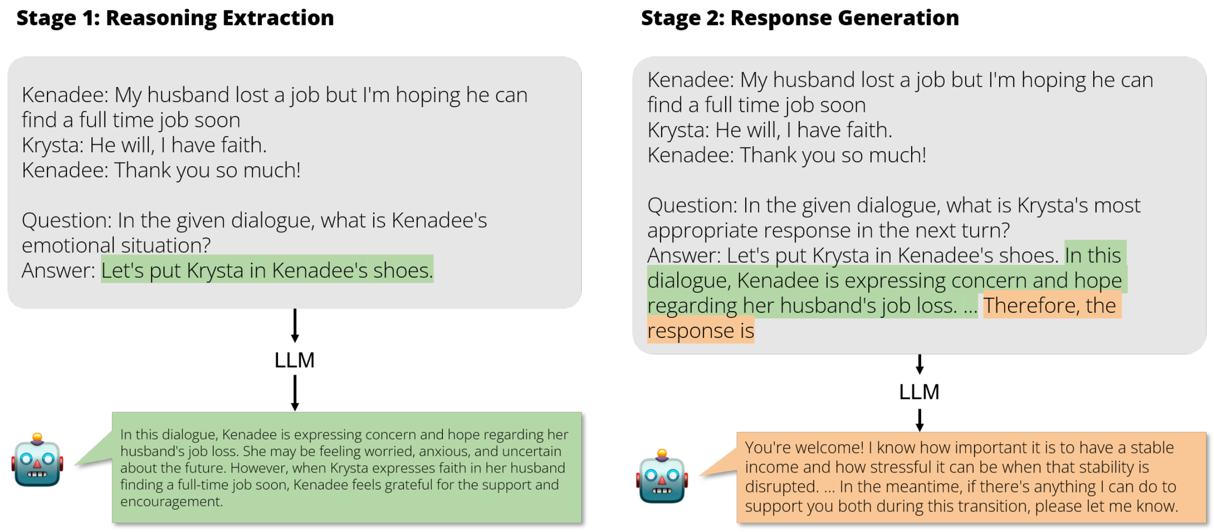


Figure 1: **An Overview of Zero-shot Perspective-Taking Prompting Method.** We present the Zero-shot Perspective-Taking prompting method, which consists of two stages: (1) Reasoning Extraction and (2) Response Generation.

Stage 1: Reasoning Extraction. This stage focuses on generating rationale R through a question-answer approach, feeding the LLM with an input prompt and a trigger sentence. The phrase "*Let's think step by step.*" is commonly used as a trigger sentence, given its proven performance boost. The Top-10K common names of US SSN applicants from 1990 to 2021 [1] are utilized to enhance naturalness in the dialogue and reduce name bias, in line with previous work (Kim et al., 2022).

Stage 2: Answer Extraction. This stage aims to generate an empathetic response y from the LLM, given the input prompt, rationale R, and another trigger sentence, "*Therefore, the response is*". Afterward, we parse the generated responses by LLM to evaluate the quality in terms of empathetic reasoning.

3.3 Zero-shot Perspective-Taking

We believe that prompting LLMs to reason in a *step-by-step* manner is ineffective for generating empathetic dialogues. This is because, unlike tasks where reasoning requires sequential consideration of the evidence for a given problem (e.g.,, arithmetic, symbolic tasks), empathetic dialogue necessitates understanding the interlocutor's emotional situation beyond the literal meaning of the given dialogue context. As such, we adopt the *perspective-taking* style (Davis, 1983; Ruby and Decety, 2004) (i.e., "*Let's put speaker into interlocutor's shoes.*") for empathetic reasoning rather than a step-by-step approach, as shown in Figure 1. Essentially, our method retains the two-stage structure of Zero-shot CoT. However, a notable difference lies in adapting the `Question` used at each stage to suit their specific objectives better. In Stage 1, we focus on a deeper understanding of the interlocutor's emotional state by posing the question "*In the given dialogue, what is Kenadee's emotional situation?*". In Stage 2, based on the grasped emotional situation, we direct LLMs to generate an appropriate follow-up empathetic response by asking "*In the given dialogue, what is Krysta's most appropriate response in the next turn?*". Indeed, in our preliminary experiments, we observed significantly better performance when using distinct questions for stages 1 and 2 instead of employing the same question for both stages.

4 Experimental Setup

4.1 Dataset

To measure the impact of Zero-shot CoT on empathetic reasoning ability, we use a test set of the EMPATHETICDIALOGUES dataset constructed via crowdsourcing to learn to express empathy adequately. This dataset consists of 25k dialogues between a speaker and a listener, where each dialogue is grounded in the emotional situation of the speaker. These situations are labeled among 32 emotion categories. In our experiment, we test

[1] `https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data`

the empathetic reasoning ability of the LLM on a subset of 1.5k dialogues sampled from the total test set of 2.5k. This setting is intended to reduce the cost of calling the OpenAI API.

4.2 Large Language Models

To explore the effect of Zero-shot CoT in the empathetic dialogue generation task, we evaluate various LLMs with different zero-shot prompting methods: Zero-shot In-Context Learning (Z-ICL), Zero-shot Chain of Thought (Z-CoT), and Zero-shot Perspective-Taking (Z-Pers). For Proprietary LLMs, we evaluate three different models: 1) variants of INSTRUCTGPT (Ouyang et al., 2022) (INSTRUCTGPT$_{d001}$, INSTRUCTGPT$_{d002}$, INSTRUCTGPT$_{d003}$), 2) CHATGPT (OpenAI, 2023a), and 3) GPT-4 (OpenAI, 2023b) [2]. For Open-Source LLMs, we evaluate 9 models in total: 1) ALPACA-13B (Taori et al., 2023), 2) DOLLY-V2-13B (Conover et al., 2023), 3) GPT-4-ALPACA-13B (Peng et al., 2023), 4) KOALA-13B (Geng et al., 2023), 5) OPENASSISTANT-13B (Köpf et al., 2023), 6) SHAREGPT-13B [3], 7) WIZARDLM-13B (Xu et al., 2023), 8) TÜLU-13B (Wang et al., 2023), and 9) LLAMA2-CHAT-13B (Touvron et al., 2023).

4.3 Implementation Details

We conduct all experiments on two A 100 (40GB) GPUs. For each stage, we set maximum tokens to 1024, temperature to 0.9, frequency penalty to 1.0, presence penalty to 0.6, top_p to 0.95, and stop tokens to `\n\n`.

4.4 Evaluation Metrics

To measure whether LLMs or dialogue generative models generate empathetic responses given the dialogue history, existing studies evaluate the generated responses on various automatic metrics related to EMPATHY. However, these evaluation metrics primarily focus on automatically evaluating the "perceived empathy" of the generated responses rather than exploring the "empathetic reasoning" abilities of LLMs. Thus, in this work, we assess how well LLM empathizes from two aspects: (1) *perceived empathy* and (2) *empathetic reasoning*.

Measuring Perceived Empathy. We evaluate the generated responses from LLMs on various metrics, EPITOME, DIFF-EPITOME, EMOACC, and INTENTACC, which is related to Empathy, followed by the prior work (Lee et al., 2022). 1) EPITOME (Sharma et al., 2020) measures the Interpretations (IP), Explorations (EX), and Emotional Reactions (ER) of the generated response by leveraging fine-tuned RoBERTa (Liu et al., 2019) model, respectively. We describe the details of EPITOME in Appendix B. 2) DIFF-EPITOME, which is a modified version of EPITOME, measures the difference IP, EX, and ER scores between the generated response and ground-truth response. 3) EMOACC measures an emotion accuracy using a fine-tuned BERT-base (Devlin et al., 2018) model on the EMPATHETICDIALOGUES dataset. 4) INTENTACC measures a response intent accuracy using a fine-tuned BERT-base model on the EMPINTENT dataset (Welivita and Pu, 2020). To mitigate the limited capacity of the previous automatic evaluation metrics, we need to conduct a more holistic and flexible evaluation method. Inspired by the recent studies where LLM-based evaluation is highly correlated with humans, we also conduct an additional evaluation using GPT-4 by measuring EMPATHY, IP, EX, ER. Specifically, given the evaluation instruction, LLM's response, and pre-defined score rubric for each metrics, GPT-4 assigns a score from 1 to 3 based on the score rubrics that have a corresponding description for each score, followed by the previous study (Ye et al., 2023). We randomly sample 120 dialogues with the same number of samples in terms of the size of dialogue history to investigate the performance variation depending on the size of dialogue history. We present detailed information in the Appendix.

Measuring Empathetic Reasoning. To measure how well LLM do empathetic reasoning, we conduct GPT-4-based evaluation by defining the scoring rubrics in terms of UNDERSTANDING and INTERPRETING based on the AEL framework.

5 Experimental Results

5.1 Does Zero-shot CoT significantly enhance empathetic understanding?

Table 1 shows the zero-shot performance of various LLMs measured regarding Empathy depending on the prompting methods (i.e., Z-ICL and Z-CoT). Zero-shot CoT generally fails to enhance the zero-shot performance of LLMs across most evaluation metrics, particularly in EMOACC, INTENTACC, and IP (refer to ▼ in Table 1). Interestingly, in the

[2] We conduct all experiments using large language models via OpenAI API from May-2023 to October-2023.

[3] https://sharegpt.com/

	EMOACC			INTENTACC			IP			EX			ER			diff-IP			diff-EX			diff-ER		
	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ
ALPACA-13B	17.33	14.13	▼ 3.20	23.93	23.33	▼ 0.60	0.25	0.04	▼ 0.21	0.48	0.88	▲ 0.39	0.86	0.84	▼ 0.01	0.89	0.72	▲ 0.17	1.17	1.75	▼ 0.59	0.86	0.80	▲ 0.06
DOLLY-V2-13B	16.67	15.13	▼ 1.54	27.53	25.87	▼ 1.66	0.39	0.29	▼ 0.10	0.12	0.18	▲ 0.06	0.70	0.86	▲ 0.16	0.99	0.91	▲ 0.08	0.69	0.79	▼ 0.10	0.76	0.89	▼ 0.13
GPT-4-ALPACA-13B	17.87	16.40	▼ 1.47	20.33	22.33	▲ 2.00	0.17	0.05	▼ 0.12	0.84	1.07	▲ 0.23	1.00	1.00	▲ 0.01	0.83	0.69	▲ 0.15	1.69	1.98	▼ 0.29	0.91	0.85	▲ 0.06
KOALA-13B	15.40	14.53	▼ 0.87	19.53	20.80	▲ 1.27	0.27	0.14	▼ 0.12	0.68	0.72	▲ 0.04	1.04	0.98	▼ 0.07	0.98	0.83	▲ 0.14	1.50	1.51	▼ 0.01	1.04	0.97	▲ 0.08
OPENASSISTANT-13B	7.20	10.00	▲ 2.80	20.80	20.60	▼ 0.20	0.07	0.08	▲ 0.01	0.61	0.53	▼ 0.08	0.57	0.64	▲ 0.06	0.78	0.77	▲ 0.01	1.41	1.31	▲ 0.10	0.83	0.89	▼ 0.05
SHAREGPT-13B	14.33	13.87	▼ 0.46	23.40	22.20	▼ 1.20	0.06	0.07	▲ 0.01	0.53	0.67	▲ 0.13	0.69	0.83	▲ 0.13	0.73	0.72	▲ 0.01	1.27	1.38	▼ 0.11	0.70	0.76	▼ 0.06
WIZARDLM-13B	16.00	17.00	▲ 1.00	22.93	19.60	▼ 3.33	0.18	0.11	▼ 0.07	0.25	0.27	▲ 0.01	0.86	1.00	▲ 0.14	0.85	0.75	▲ 0.09	0.87	0.89	▼ 0.02	0.85	0.88	▼ 0.03
TÜLU-13B	16.13	16.33	▲ 0.20	23.67	22.20	▼ 1.47	0.10	0.11	▲ 0.01	0.91	0.57	▼ 0.34	0.78	1.00	▲ 0.22	0.73	0.79	▼ 0.06	1.73	1.27	▲ 0.46	0.75	0.94	▼ 0.19
LLAMA2-CHAT-13B	14.40	5.40	▼ 9.00	21.47	13.87	▼ 7.60	0.08	0.02	▼ 0.07	0.65	0.24	▼ 0.41	1.17	0.33	▼ 0.84	0.76	0.71	▲ 0.05	1.42	0.93	▲ 0.49	0.98	0.71	▲ 0.27
INSTRUCTGPT$_{d001}$	16.33	12.27	▼ 4.06	27.07	22.60	▼ 4.47	0.26	0.15	▼ 0.11	0.41	0.33	▼ 0.08	0.84	0.41	▼ 0.42	0.85	0.81	▲ 0.04	1.03	0.97	▲ 0.06	0.78	0.74	▲ 0.04
INSTRUCTGPT$_{d002}$	15.13	12.07	▼ 3.06	27.40	20.27	▼ 7.13	0.28	0.15	▼ 0.12	0.25	0.41	▲ 0.16	0.98	0.62	▼ 0.36	0.87	0.77	▲ 0.10	0.85	1.09	▼ 0.24	0.93	0.89	▲ 0.04
INSTRUCTGPT$_{d003}$	18.00	16.80	▼ 1.20	26.47	22.60	▼ 3.87	0.13	0.05	▼ 0.08	0.98	0.94	▼ 0.04	1.11	0.79	▼ 0.32	0.76	0.70	▲ 0.06	1.81	1.85	▼ 0.04	0.96	0.82	▲ 0.15
CHATGPT	17.93	18.67	▲ 0.74	25.93	22.40	▼ 3.53	0.16	0.07	▼ 0.08	0.40	0.47	▲ 0.07	0.98	1.05	▲ 0.07	0.76	0.74	▲ 0.03	1.08	1.18	▼ 0.10	0.73	0.83	▼ 0.10
GPT-4	18.93	19.20	▲ 0.27	27.07	23.40	▼ 3.67	0.08	0.04	▼ 0.04	0.70	0.64	▼ 0.05	1.03	1.08	▲ 0.05	0.75	0.67	▲ 0.08	1.42	1.35	▲ 0.07	0.78	0.83	▼ 0.05

Table 1: **Zero-shot Performance between Z-ICL and Z-CoT on Perceived Empathy.** We evaluate the zero-shot performance of LLMs using different prompting methods (i.e., Z-ICL and Z-CoT) across various metrics (i.e., EMOACC, INTENTACC, EPITOME, and DIFF-EPITOME) for measuring the perceived empathy performance using automatic metrics. ▼ and ▲ indicate a performance decrease and increase when Z-CoT is applied, respectively. Δ represents the difference in performance between Z-CoT and Z-ICL (i.e., Z-CoT - Z-ICL).

	EMPATHY			IP			EX			ER			diff-IP			diff-EX			diff-ER		
Model	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ	Z-ICL	Z-CoT	Δ
ALPACA-13B	2.34	2.56	▲ 0.22	2.15	2.3	▲ 0.15	1.39	1.68	▲ 0.29	1.94	1.98	▲ 0.04	0.87	0.93	▼ 0.06	0.55	0.96	▼ 0.41	1.37	1.37	(=) 0
DOLLY-V2-13B	2.12	2.2	▲ 0.08	1.96	2.03	▲ 0.07	1.15	1.13	▼ 0.02	1.6	1.71	▲ 0.11	0.92	0.84	▲ 0.08	0.35	0.29	▲ 0.06	1.12	1.27	▼ 0.15
GPT-4-ALPACA-13B	2.72	2.69	▼ 0.03	2.66	2.58	▼ 0.08	1.91	1.84	▼ 0.07	2.41	2.47	▲ 0.06	1.71	1.52	▲ 0.19	1.41	1.34	▲ 0.07	2.17	2.33	▼ 0.16
KOALA-13B	2.37	2.38	▲ 0.01	2.21	2.17	▼ 0.04	1.58	1.51	▼ 0.07	1.98	1.99	▲ 0.01	1.2	1.18	▲ 0.02	1	0.89	▲ 0.11	1.47	1.66	▼ 0.19
OPENASSISTANT-13B	1.47	1.79	▲ 0.32	1.42	1.73	▲ 0.31	1.04	1.33	▲ 0.29	1.13	1.59	▲ 0.46	1.19	1.25	▼ 0.06	0.38	0.61	▼ 0.23	0.86	1.46	▼ 0.6
SHAREGPT-13B	2.29	2.35	▲ 0.06	2.27	2.27	(=) 0	1.46	1.66	▲ 0.2	1.83	2.12	▲ 0.29	1.35	1.7	▼ 0.35	0.7	1.06	▼ 0.36	1.37	2.11	▼ 0.74
WIZARDLM-13B	2.52	2.69	▲ 0.17	2.5	2.69	▲ 0.19	1.54	1.9	▲ 0.36	2.17	2.58	▲ 0.41	1.73	1.76	▼ 0.03	0.91	1.4	▼ 0.49	2.03	2.56	▼ 0.53
TÜLU-13B	2.52	2.46	▼ 0.06	2.42	2.25	▼ 0.17	1.7	1.51	▼ 0.19	2.03	2.04	▲ 0.01	1.52	1.17	▲ 0.35	1.17	0.72	▲ 0.45	1.53	1.54	▼ 0.01
LLAMA2-CHAT-13B	2.41	1.07	▼ 1.34	2.34	1.07	▼ 1.27	1.77	1.01	▼ 0.76	2.27	1.03	▼ 1.24	1.69	1.49	▲ 0.2	1.3	0.64	▲ 0.66	2.25	1	▲ 1.25
INSTRUCTGPT$_{d001}$	2.24	1.33	▼ 0.91	2.12	1.21	▼ 0.91	1.33	1.02	▼ 0.31	1.71	1.14	▼ 0.57	0.86	0.89	▼ 0.03	0.43	0.25	▲ 0.18	0.96	0.65	▲ 0.31
INSTRUCTGPT$_{d002}$	2.34	1.6	▼ 0.74	2.16	1.45	▼ 0.71	1.23	1.08	▼ 0.15	1.84	1.27	▼ 0.57	0.89	0.99	▼ 0.1	0.38	0.31	▲ 0.07	1.14	0.87	▲ 0.27
INSTRUCTGPT$_{d003}$	2.87	2.26	▼ 0.61	2.62	2.09	▼ 0.53	1.56	1.39	▼ 0.17	2.41	1.72	▼ 0.69	1.28	0.92	▲ 0.36	0.81	0.61	▲ 0.2	1.92	1.08	▲ 0.84
CHATGPT	2.92	2.92	(=) 0	2.86	2.88	▲ 0.02	1.58	1.65	▲ 0.07	2.5	2.61	▲ 0.11	1.79	1.88	▼ 0.09	0.78	0.97	▼ 0.19	2.15	2.34	▼ 0.19
GPT-4	2.98	2.99	▲ 0.01	2.91	2.95	▲ 0.04	1.74	1.98	▲ 0.24	2.78	2.85	▲ 0.07	1.82	1.95	▼ 0.13	0.92	1.24	▼ 0.32	2.65	2.85	▼ 0.2

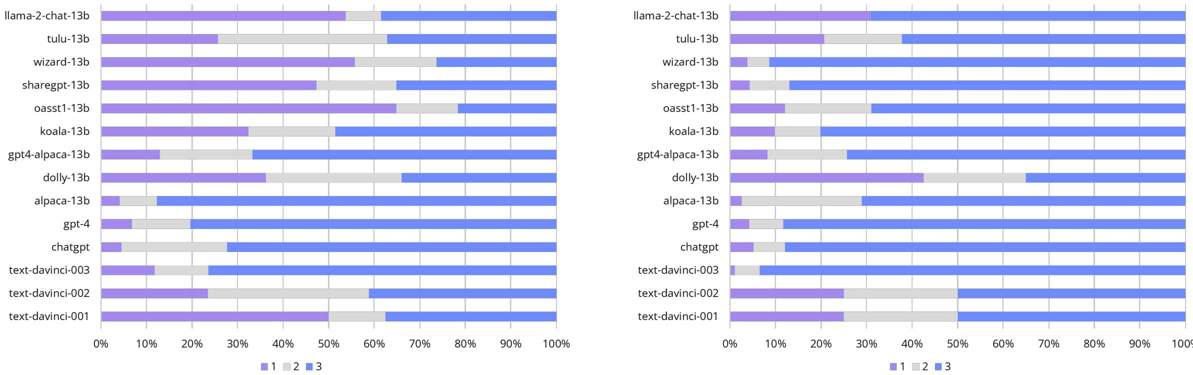
Table 2: **Zero-shot Performance between Z-ICL and Z-CoT on Perceived Empathy using GPT-4-based Evaluation.** We evaluate the zero-shot performance of LLMs using different prompting methods (i.e., Z-ICL and Z-CoT) for measuring the perceived empathy performance. Δ represents the difference in performance between Z-CoT and Z-ICL (i.e., Z-CoT - Z-ICL).

IP metrics, Z-CoT do not yield favorable results. However, in the diff-IP metrics, the application of Z-CoT improves the performance of all LLMs. This underscores that prompting the model to *think step-by-step* aids the "Processing" component of AEL framework, allowing us to emulate human empathetic patterns in genuine empathy-based conversations. On the other hand, in the diff-EX metrics, Z-CoT fails to enhance overall performance, indicating that language models are excessively curious about the current situation while increasing their understanding of the current situation due to rationale. In particular, for open-sourced LLMs, the performance consistently decreases except for OPENASSISTANT-13B.

While evaluating responses from general LLMs is challenging due to automatic evaluation metrics (i.e., BERT and RoBERTa) fine-tuned for measuring empathic performance (refer to Limitations), we assess EMPATHY and EPITOME for 120 samples using GPT-4, with results in Table x. This exhibits a different trend compared to Table 1. Notably, open-sourced LLMs tend to show improved EMPATHY performance, potentially benefiting from the Z-CoT prompting method. Yet, despite ER improvements, most LLMs (excluding LLAMA2-CHAT-13B) under the diff-ER metric show a marked decline, indicating an overproduction of emotional expressions. For Proprietary LLMs, trends align with Table 1. Unlike their open-sourced counterparts, they don't seem to benefit significantly from Z-CoT. There's a noticeable decline in the EMPATHY metric, raising questions about Z-CoT's efficacy in social knowledge reasoning versus other symbolic reasoning tasks. This mirrors prior findings on reduced performance in IP and diff-IP associated with "Processing". Interestingly, all LLMs show patterns where EPITOME and

Model	EMPATHY			IP			EX			ER			diff-IP			diff-EX			diff-ER		
	CoT	Pers	Δ	CoT	Pers	Δ	CoT	Pers	Δ	CoT	Pers	Δ	CoT	Pers	Δ	CoT	Pers	Δ	CoT	Pers	Δ
ALPACA-13B	2.56	2.6	▲ 0.04	2.3	2.33	▲ 0.03	1.68	1.43	▼ 0.25	1.98	2.17	▲ 0.19	0.93	1.04	▲ 0.11	0.96	0.6	▼ 0.36	1.37	1.6	▲ 0.23
DOLLY-V2-13B	2.2	1.99	▼ 0.21	2.03	1.83	▼ 0.2	1.13	1.14	▲ 0.01	1.71	1.58	▼ 0.13	0.84	0.93	▲ 0.09	0.29	0.29	(=) 0	1.27	1.11	▼ 0.16
GPT-4-ALPACA-13B	2.69	2.75	▲ 0.06	2.58	2.6	▲ 0.02	1.84	1.81	▼ 0.03	2.47	2.42	▼ 0.05	1.52	1.47	▼ 0.05	1.34	1.27	▼ 0.07	2.33	2.06	▼ 0.27
KOALA-13B	2.38	2.53	▲ 0.15	2.17	2.38	▲ 0.21	1.51	1.62	▲ 0.11	1.99	2.1	▲ 0.11	1.18	1.21	▲ 0.03	0.89	0.92	▲ 0.03	1.66	1.65	▼ 0.01
OPENASSISTANT-13B	1.79	2.14	▲ 0.35	1.73	2.01	▲ 0.28	1.33	1.47	▲ 0.14	1.59	1.91	▲ 0.32	1.25	1.23	▼ 0.02	0.61	0.77	▲ 0.16	1.46	1.56	▲ 0.1
SHAREGPT-13B	2.35	2.62	▲ 0.27	2.27	2.47	▲ 0.2	1.66	1.73	▲ 0.07	2.12	2.45	▲ 0.33	1.7	1.52	▼ 0.18	1.06	0.98	▼ 0.08	2.11	2.3	▲ 0.19
WIZARDLM-13B	2.69	2.76	▲ 0.07	2.69	2.73	▲ 0.04	1.9	1.92	▲ 0.02	2.58	2.64	▲ 0.06	1.76	1.87	▲ 0.11	1.4	1.28	▼ 0.12	2.56	2.63	▲ 0.07
TÜLU-13B	2.46	2.61	▲ 0.15	2.25	2.46	▲ 0.21	1.51	1.61	▲ 0.1	2.04	2.32	▲ 0.28	1.17	1.41	▲ 0.24	0.72	0.89	▲ 0.17	1.54	2.03	▲ 0.49
LLAMA2-CHAT-13B	1.07	1.42	▲ 0.35	1.07	1.43	▲ 0.36	1.01	1.28	▲ 0.27	1.03	1.38	▲ 0.35	1.49	1.33	▼ 0.16	0.64	0.83	▲ 0.19	1	1.25	▲ 0.25
INSTRUCTGPT$_{d001}$	1.33	1.51	▲ 0.18	1.21	1.35	▲ 0.14	1.02	1.04	▲ 0.02	1.14	1.2	▲ 0.06	0.89	0.8	▼ 0.09	0.25	0.2	▼ 0.05	0.65	0.69	▲ 0.04
INSTRUCTGPT$_{d002}$	1.6	1.59	▼ 0.01	1.45	1.45	(=) 0	1.08	1.04	▼ 0.04	1.27	1.26	▼ 0.01	0.99	0.83	▼ 0.16	0.31	0.25	▼ 0.06	0.87	0.7	▼ 0.17
INSTRUCTGPT$_{d003}$	2.26	2.73	▲ 0.47	2.09	2.5	▲ 0.41	1.39	1.35	▼ 0.04	1.72	2.22	▲ 0.5	0.92	1.3	▲ 0.38	0.61	0.57	▼ 0.04	1.08	1.6	▲ 0.52
CHATGPT	2.92	2.98	▲ 0.06	2.88	2.89	▲ 0.01	1.65	1.68	▲ 0.03	2.61	2.63	▲ 0.02	1.88	1.86	▼ 0.02	0.97	0.89	▼ 0.08	2.34	2.37	▲ 0.03
GPT-4	2.99	3	▲ 0.01	2.95	2.94	▼ 0.01	1.98	1.98	(=) 0	2.85	2.89	▲ 0.04	1.95	1.98	▲ 0.03	1.24	1.29	▲ 0.05	2.85	2.91	▲ 0.06

Table 3: **Zero-shot Performance between Z-CoT and Z-Pers of Perceived Empathy using GPT-4-based Evaluation.** We evaluate the zero-shot performance of LLMs using different prompting methods (i.e., Z-CoT and Z-Pers) on various metrics using GPT-4-based evaluation for measuring the perceived empathy performance. Δ represents the difference in performance between Z-Pers and Z-CoT (i.e., Z-Pers - Z-CoT).



(a) Z-CoT

(b) Z-Pers

Figure 2: **Percent Distribution of Scores on INTERPRETING when EMPATHY = 3.** We show the distribution percentages for INTERPRETING when the LLM achieves an EMPATHY score of 3, indicating a high score. From this visualization, two key observations can be made: (1) Z-Pers boost perceived empathy performance, and (2) Greater empathetic reasoning is essential for generating superior empathetic responses.

DIFF-EPITOME metrics are inversely related. Such observations highlight the need for comprehensive evaluation metrics tailored for the empathetic dialogue generation task.

Rather than solely assessing the empathetic capability from LLM-generated responses, we evaluate the model's genuine aptitude for empathetic reasoning. In the UNDERSTANDING, GPT-4exhibits the most proficient understanding of the interlocutor's emotional situation, followed closely by the ALPACA-13B model. Among Open-sourced LLMs, LLAMA2-CHAT-13B and OPENASSISTANT-13B underperformed, given their specialization in dialogue safety, helpfulness, and their training through forward modeling. This implies that crafting safe and helpful responses and excelling in empathy necessitate distinct capabilities. With Proprietary LLMs, there was a consistent performance uptrend with model advancements. Notably, scores in the INTERPRETING are generally lower than those in the UNDERSTANDING metric, highlighting the increased complexity of discerning and interpreting implicit meanings in interlocutor's utterances beyond mere literal comprehension. Even when leveraging Z-CoT, the achieved performance metrics are not particularly impressive, suggesting that Z-CoT might not be an optimal solution for the empathetic dialogue generation task.

Model	UNDERSTANDING Z-CoT	Z-Pers	Δ	INTERPRETING Z-CoT	Z-Pers	Δ
ALPACA-13B	2.92	2.9	▼ 0.02	2.78	2.65	▼ 0.13
DOLLY-V2-13B	2.34	2.27	▼ 0.07	1.85	1.89	▲ 0.04
GPT-4-ALPACA-13B	2.7	2.81	▲ 0.11	2.44	2.59	▲ 0.15
KOALA-13B	2.53	2.71	▲ 0.18	2.11	2.61	▲ 0.5
OPENASSISTANT-13B	1.41	2.58	▲ 1.17	1.3	2.35	▲ 1.05
SHAREGPT-13B	2.02	2.89	▲ 0.87	1.78	2.77	▲ 0.99
WIZARDLM-13B	1.95	2.89	▲ 0.94	1.65	2.85	▲ 1.2
TÜLU-13B	2.48	2.59	▲ 0.11	2.01	2.39	▲ 0.38
LLAMA2-CHAT-13B	1.75	2.05	▲ 0.3	1.64	1.98	▲ 0.34
INSTRUCTGPT$_{d001}$	2.02	2.17	▲ 0.15	1.79	1.92	▲ 0.13
INSTRUCTGPT$_{d002}$	2.42	2.03	▼ 0.39	2.13	1.86	▼ 0.27
INSTRUCTGPT$_{d003}$	2.53	2.91	▲ 0.38	2.33	2.83	▲ 0.5
CHATGPT	2.89	2.93	▲ 0.04	2.62	2.83	▲ 0.21
GPT-4	2.95	2.99	▲ 0.04	2.72	2.84	▲ 0.12

Table 4: **Zero-shot Performance of Empathetic Reasoning.** We evaluate the zero-shot performance of LLMs using different prompting methods (i.e., Z-CoT and Z-Pers) on UNDERSTANDING and INTERPRETING for measuring the empathetic reasoning capability. Δ represents the difference in performance between Z-Pers and Z-CoT (i.e., Z-Pers - Z-CoT).

# of Prev. Utter →	3			5			7		
Model	Z CoT	Z Pers	Δ	Z CoT	Z Pers	Δ	Z CoT	Z Pers	Δ
ALPACA-13B	2.78	2.58	▼ 0.2	2.7	2.65	▼ 0.05	2.85	2.71	▼ 0.14
DOLLY-V2-13B	1.91	1.71	▼ 0.2	1.91	1.96	▲ 0.05	1.76	1.99	▲ 0.23
GPT-4-ALPACA-13B	2.4	2.49	▲ 0.09	2.44	2.44	(=) 0	2.48	2.84	▲ 0.36
KOALA-13B	2.08	2.58	▲ 0.5	2.25	2.61	▲ 0.36	2	2.65	▲ 0.65
OPENASSISTANT-13B	1.35	2.3	▲ 0.95	1.29	2.34	▲ 1.05	1.27	2.41	▲ 1.14
SHAREGPT-13B	2.02	2.58	▲ 0.56	1.74	2.9	▲ 1.16	1.53	2.84	▲ 1.31
WIZARDLM-13B	1.75	2.81	▲ 1.06	1.48	2.85	▲ 1.37	1.74	2.88	▲ 1.14
TÜLU-13B	1.8	2.22	▲ 0.42	2.06	2.44	▲ 0.38	2.17	2.5	▲ 0.33
LLAMA2-CHAT-13B	2.08	2.08	(=) 0	1.54	2	▲ 0.46	1.31	1.82	▲ 0.51
INSTRUCTGPT$_{d001}$	1.81	1.93	▲ 0.12	1.95	1.91	▼ 0.04	1.63	1.88	▲ 0.25
INSTRUCTGPT$_{d002}$	2.31	1.76	▼ 0.55	2.06	1.65	▼ 0.41	2.01	2.12	▲ 0.11
INSTRUCTGPT$_{d003}$	2.15	2.82	▲ 0.67	2.52	2.85	▲ 0.33	2.31	2.82	▲ 0.51
CHATGPT	2.55	2.65	▲ 0.1	2.68	2.94	▲ 0.26	2.64	2.9	▲ 0.26
GPT-4	2.52	2.7	▲ 0.18	2.79	2.85	▲ 0.06	2.86	2.98	▲ 0.12

Table 5: **Zero-shot Performance Depending on Dialogue History.** We show the variation of zero-shot performance depending on the number of dialogue history. # of Prev. Utter denotes the number of utterances of a given dialogue.

5.2 *Perspective-Taking* can improve empathetic reasoning

As shown in Table 4, Z-CoT doesn't seem to be the optimal prompting method for extracting the empathetic reasoning abilities of LLM, especially when considering the performance of INTERPRETING. This could be attributed to Z-CoT's tendency to infer more from the literal meaning of given dialogues, often at the expense of a deeper understanding. In empathy, the perspective-taking process, which involves interpreting situations from the client's viewpoint, is of paramount importance. Therefore, we investigate whether Z-Pers, a prompt designed to encourage LLMs to engage in the *perspective-taking* process, could enhance empathetic reasoning performance. Table 4 shows that Z-Pers improves performance in both UNDERSTANDING and INTERPRETING metrics. This suggests that, much like humans, LLMs benefit significantly from the induction of the perspective-taking process and inherently possess some capacity for it. For models such as ALPACA-13B and INSTRUCTGPT$_{d002}$, there are instances where responses are generated directly without prior reasoning, indicating a potential evaluation that the scenario doesn't warrant a deeper view from the interlocutor's perspective.

Enhancing empathetic reasoning also boosts perceived empathy. We investigate whether perspective-taking enhances empathetic reasoning and perceived empathy performance. Due to the limitations of automatic evaluation metrics, we leverage GPT-4for the assessment. As depicted in Table 3, Z-Pers consistently outperforms Z-CoT in the EMPATHY performance across most LLMs. This suggests that while empathetic reasoning is essential for generating empathetic responses, a perspective-taking style may be more effective than a step-by-step method. Moreover, we observe performance enhancements in both the EPITOME and DIFF-EPITOME metrics when using Z-Pers. Notably, while the introduction of perspective-taking lead to a remarkable increase in IP scores (as opposed to the decrease associated with Z-CoT in Tables 1 and 2), the diff-IP performance generally declined for most LLMs with the implementation of Z-Pers. This indicates that while extensively interpreting situations from the interlocutor's viewpoint - thus elevating the IP score - Z-Pers diverges from typical human patterns of interpreting the interlocutor's emotional situation. It's clear that humans do not continually employ perspective-taking throughout a conversation; sometimes, they opt for more superficial empathetic expressions such as "I'm sorry to hear that" or "I'm so happy for you." This underscores the importance of balancing perceived empathy with empathetic reasoning, contingent on the specific context.

We investigate the correlation between EMPATHY and empathetic reasoning based on each LLM's scores obtained from GPT-4-based evalua-

Dialogue:

Kenadee: My husband lost a job but I'm hoping he can find a full time job soon

Krysta: He will, I have faith.

Kenadee: Thank you so much!

Golden Response: No problem. What kind of work does he do?

Z-ICL:

[Response]: You're welcome, Kenadee. I know how stressful it can be when someone loses their job, but I truly believe that your husband will find a full-time job soon.

Z-CoT:

[Rationale]: Kenadee: My husband lost a job but I'm hoping he can find a full-time job soon.

From this statement, it is clear that Kenadee is feeling concerned and worried about her husband losing his job. She may be experiencing stress and uncertainty about their financial situation.

[Response]: I understand how stressful this situation can be. It's completely understandable to feel worried about your husband's job loss.

Z-Pers:

[Rationale]: In this dialogue, Kenadee is expressing concern and hope regarding her husband's job loss. She may be feeling worried, anxious, and uncertain about the future. However, when Krysta expresses faith in her husband finding a full-time job soon, Kenadee feels grateful for the support and encouragement.

[Response]: You're welcome! I know how important it is to have a stable income and how stressful it can be when that stability is disrupted. But remember, your husband has skills and experience that will make him an asset to any employer. It might take some time, but I truly believe he will find a full-time job soon. In the meantime, if there's anything I can do to support you both during this transition, please let me know.

Table 6: **Case Study.** We present results generated by GPT-4with the Zero-shot ICL, Zero-shot CoT, and Zero-shot Pers prompt methods for a given dialogue.

tion. Figure 2 illustrates the statistics of the scores that each LLM received for empathetic reasoning in INTERPRETING when the EMPATHY value was 3 points (high score). As evident from Figure 2, when the LLM attempts high empathy, Z-Pers induces deeper empathetic reasoning than Z-CoT. Furthermore, the proportion that achieves a score of 1 point decreases drastically. This confirms that Z-Pers perform better in EMPATHY (refer to Table 4) while attempting more enhanced empathetic reasoning.

Z-Pers is more effective with extended dialogue history. Table 5 shows the variation in INTERPRETING performance based on the number of utterances provided in a dialogue history for generating empathetic responses. Overall, Z-Pers consistently outperforms Z-CoT in enhancing the LLM's ability to interpret the interlocutor's emotional state, regardless of the utterance count. Notably, as the number of utterances in the dialogue history increases, Z-Pers enables the LLM to achieve a more nuanced understanding, leveraging the accumulating hints (e.g., feelings, thoughts, emotions, key entities) about the interlocutor's emotional situation.

Case Study. Table 6 presents generated responses and rationales by CHATGPTusing different prompting methods (i.e., Z-ICL, Z-CoT, and Z-Pers) in a zero-shot setting. All three different prompting methods (i.e., Z-ICL, Z-CoT, and Z-Pers) can generate appropriate empathetic responses to the interlocutor. Compared to the golden response, the generated responses are generally longer, a trend observed in LLMs trained with Reinforcement Learning with Human Feedback (RLHF). Among them, Z-Pers produce the longest responses. This is because Z-Pers induce an additional step in the perspective-taking process to understand the interlocutor's emotional situation from their point of view, leading to deeper interpretations in the generated responses. Z-CoT focuses on analyzing the literal meanings of the given conversation context, resulting in simple analyses of the present situation. As shown in Figure 6, with Z-Pers, when Krysta expresses "faith," it can infer how Kenadee might feel from Kenadee's perspective. Consequently, it generates responses that evoke deeper empathy.

6 Conclusion

This work investigates the effectiveness of Zero-shot CoT in enhancing the empathetic reasoning capability of LLM. Our experiments reveal that Zero-shot CoT does not improve zero-shot performance in the empathetic dialogue generation task on various metrics. The *perspective-taking* prompting method leads to improved performance on both the perceived empathy and empathetic reasoning compared to Zero-shot CoT. In future research, we plan to introduce a pragmatic-reasoning-based prompting method and comprehensive, robust evaluation metrics for assessing the empathetic reasoning abilities of LLMs.

Limitations

Limited Capacity of Automatic Metrics. As is well known, empathy is an exceptionally subjective characteristic. Therefore, assessing it can be quite challenging, as individuals may perceive different degrees of empathy. Although many studies (Sharma et al., 2020; Kim et al., 2021; Lee et al., 2022) have proposed various metrics for empathetic reasoning, there are three limitations in quantitatively evaluating the empathetic reasoning capability of LLMs. 1) Many evaluation metrics are machine-based methods, fine-tuning models like BERT (Devlin et al., 2018). These metrics can cause inaccurate performance and are insufficient to evaluate the diverse and high-quality responses generated by InstructGPT (Ouyang et al., 2022). Recently proposed prompt-based evaluators (Fu et al., 2023; Liu et al., 2023) might help but can also prefer the LLM-generated responses, reported in a prior work (Liu et al., 2023). 2) The datasets used for machine-based evaluation are somewhat limited in their domain and dialogue diversity. For example, EPITOME and DIFF-EPITOME utilize mental health support dataset, which do not represent true open-domain social dialogue. Similarly, EMOACC and INTENTACC, which use the emotion and intent-annotated EMPATHETICDIALOGUES dataset, might fail to deliver trustworthy evaluations for responses that are uncommon in the EMPATHETICDIALOGUES dataset. 3) Current evaluation metrics evaluate empathy individually, based on different criteria. To facilitate a fairer comparison of language models in the future, a holistic, universal metric is needed to encapsulate all aspects of empathy. Considering these three limitations, there is a need for the future development of more robust and universal evaluation methods for empathetic dialogue generation task.

Lack of Pragmatic-based Prompting. As *perspective-taking* is essential in empathetic reasoning (as proven by a perspective-taking prompt), it is important for LLMs to understand the interlocutor's emotional situation. However, our experiments suggest that even the popular prompting method (i.e., Zero-shot CoT), though successful in logical reasoning tasks, is not specifically designed for empathetic reasoning (i.e., *Processing* dimension of AEL). To enhance the empathetic reasoning capability of LLM, it is necessary to develop a new prompting method incorporating *pragmatic reasoning*, enabling LLM to infer the implications of the interlocutor's messages. In a recent study (Kim et al., 2021), the RSA framework (Frank and Goodman, 2012) has previously been used to show an increase in empathetic dialogue generation across various dialogue generative models, such as MIME (Majumder et al., 2020), DodecaTransformer (Shuster et al., 2019), and Blender (Roller et al., 2020). Thus, given the importance of pragmatic reasoning, we believe that the pragmatic reasoning-based prompting method will unlock the empathetic reasoning and theory-of-mind (ToM) capabilities of LLMs.

Lack of Human Evaluation. To accurately analyze the effects of Zero-shot CoT and Zero-shot Perspective-Taking in empathetic dialogue generation, it is necessary to evaluate performance through human evaluation, considering different personalities (e.g., MBTI or Big Five Personality Traits). We will address this in future work.

References

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

Stephen B Castleberry and C David Shepherd. 1993. Effective interpersonal listening and personal selling. *Journal of Personal Selling & Sales Management*, 13(1):35–49.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Lucette B Comer and Tanya Drollinger. 1999. Active empathetic listening and selling success: A conceptual framework. *Journal of Personal Selling & Sales Management*, 19(1):15–29.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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

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

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. *arXiv preprint arXiv:2109.08828*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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

Jeremy Main. 1985. How to sell by listening. *Fortune*, 111(3):52–54.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Shima Rahimi Moghaddam and Christopher J Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

OpenAI. 2023a. ChatGPT. `https://openai.com/blog/chatgpt/`.

OpenAI. 2023b. Gpt-4 technical report. *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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

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

Perrine Ruby and Jean Decety. 2004. How would you feel versus how do you think she would feel? a neuroimaging study of perspective-taking with social emotions. *Journal of cognitive neuroscience*, 16(6):988–999.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2019. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *arXiv preprint arXiv:1911.03768*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022. Care: Causality reasoning for empathetic responses by conditional graph generation. *arXiv preprint arXiv:2211.00255*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Dialogue:

Jaylen: My cat was playing last nigh and brought it to me.

Demetrius: That sounds like fun. What did it bring to you?

Jaylen: Well, it wasn't one of his toys like I thought. It was a dead mouse

Golden Response: Eww... That's pretty gross. I had a cat once that would bring me dead birds too. I think that means they like you!

Z-ICL:

[Response]: Oh no! That must have been really scary for you.

Z-CoT:

[Rationale]: The first thing Demetrius should do is identify Jaylen's feelings. In this case, Jaylen is most likely feeling grossed out and maybe a little surprised. Demetrius should then respond in a way that shows he empathizes with these feelings. The most appropriate response would be something along the lines of "That sounds really unpleasant. I'm sorry you had to go through that."

[Response]: B.

Z-Pers:

[Rationale]: Jaylen's cat brought him a dead mouse, which grossed him out. The most appropriate response would be something like, "Ewww, that's gross! I'm sorry that happened."

[Response]: Ewww, that's gross! I'm sorry that happened.

Table 7: **Case Study.** We present results generated by CHATGPTwith the Zero-shot ICL, Zero-shot CoT, and Zero-shot Perspective-Taking prompt methods for a given dialogue.

A Prompt Template

As shown in Figure 3, we present prompt templates of Zero-shot ICL, Zero-shot CoT, and Zero-shot Perspective-Taking used in our experiments.

B EPITOME-based Automatic Evaluation

EPITOME, introduced by (Sharma et al., 2020), is a new conceptual framework for expressing empathy in text-based, asynchronous contexts. EPITOME consists of three communication mechanisms of empathy: **Explorations** (EX), **Interpretations** (IP), and **Emotional Reactions** (ER). The mechanisms are described as follows:

- EXPLORATIONS (EX) are expressions of active interest in the interlocutor's situation.
- INTERPRETATIONS (IP) are expressions of acknowledgments or understanding of the interlocutor's emotion or situation.
- EMOTIONAL REACTIONS (ER) are expressions of emotions such as warmth, compassion, and concern in the interlocutor's situation.

In a recent study (Kim et al., 2021), each mechanism was used as an automatic metric to measure the empathy of generated responses using a fine-tuned RoBERTa (Liu et al., 2019) model. Each generated response was measured by one of the values (0, 1, or 2) predicted from the model.

C More Examples

We present more generated examples with different prompting methods, as shown in Table 7 and 8.

Dialogue:

Tevin: Hello the plants in my garden I water them, I move the earth, I try to keep them happy

Vedansh: Gardening can be peaceful

Tevin: Especially when you feel something special by nature

Golden Response: Nature is a wonderful thing

Z-ICL:

[Response]: Yes, I can imagine that.

Z-CoT:

[Rationale]: Tevin has said that he enjoys gardening because it makes him feel close to nature. The most appropriate response from Vedansh would be to say something along the lines of "I know how you feel. I also enjoy spending time in my garden."

[Response]: Gardening can be peaceful.

Z-Pers:

[Rationale]: Vedansh would most likely respond with, "I know how you feel. I love spending time in my garden too."

[Response]: I know how you feel. I love spending time in my garden too.

Table 8: **Case Study 2.** We present results generated by CHATGPTwith the Zero-shot ICL, Zero-shot CoT, and Zero-shot Perspective-Taking prompt methods for a given dialogue.

Prompt Template for Zero-shot ICL:
The following dialogue is between Demetrius and Jaylen. Imagine you are Demetrius, and you should empathize well with Jaylen's situation, feelings, and thoughts. The dialogue is provided line-by-line.

Dialogue:
Jaylen: My cat was playing last nigh and brought it to me.
Demetrius: That sounds like fun. What did it bring to you?
Jaylen: Well, it wasn't one of his toys like I thought. It was a dead mouse
Demetrius:

Prompt Template for Zero-shot CoT:
The following dialogue is between Demetrius and Jaylen. Imagine you are Demetrius, and you should empathize well with Jaylen's situation, feelings, and thoughts. The dialogue is provided line-by-line.

Dialogue:
Jaylen: My cat was playing last nigh and brought it to me.
Demetrius: That sounds like fun. What did it bring to you?
Jaylen: Well, it wasn't one of his toys like I thought. It was a dead mouse

Question: In the given dialogue, what is the most appropriate response?
Answer: Let's think step by step.

Prompt Template for Zero-shot Perspective-Taking (Stage 1):
The following dialogue is between Demetrius and Jaylen. Imagine you are Demetrius, and you should empathize well with Jaylen's situation, feelings, and thoughts. The dialogue is provided line-by-line.

Dialogue:
Jaylen: My cat was playing last nigh and brought it to me.
Demetrius: That sounds like fun. What did it bring to you?
Jaylen: Well, it wasn't one of his toys like I thought. It was a dead mouse

Question: In the given dialogue, what is Jaylen's emotional situation?
Answer: Let's put Demetrius in Jaylen's shoes.

Prompt Template for Zero-shot Perspective-Taking (Stage 2):
The following dialogue is between Demetrius and Jaylen. Imagine you are Demetrius, and you should empathize well with Jaylen's situation, feelings, and thoughts. The dialogue is provided line-by-line.

Dialogue:
Jaylen: My cat was playing last nigh and brought it to me.
Demetrius: That sounds like fun. What did it bring to you?
Jaylen: Well, it wasn't one of his toys like I thought. It was a dead mouse

Question: In the given dialogue, what is the most appropriate response in the next turn?
Answer: Let's put Demetrius in Jaylen's shoes. `[Rationale]` Therefore, the response is

Figure 3: **Prompt Templates.** A prompt template for Zero-shot ICL (**top**). A prompt template for Zero-shot CoT (**middle**). A prompt template for Zero-shot Perspective-Taking (**bottom**).