Investigating the Effects of Zero-Shot Chain-of-Thought on Empathetic Dialogue Generation

Young-Jun Lee[1] **Dokyong Lee**[2] **Jihui Im**[2] **Joo Won Sung**[2] **Ho-Jin Choi**[1]
[1]School of Computing, KAIST [2]KT Corporation
{yj2961, hojinc}@kaist.ac.kr {dokyong.lee, jihui.im, jwsung}@kt.com

Abstract

This study investigates the effectiveness of the Zero-shot Chain-of-Thought (CoT) approach, specifically the "*Let's think step by step.*", in boosting the empathetic reasoning capabilities of Large Language Models (LLMs). Our experiments, however, reveal that Zero-shot CoT does not sufficiently enhance the empathetic reasoning of LLMs as compared to Zero-shot In-Context Learning (ICL), according to a variety of performance metrics. Importantly, we discovered that the perspective-taking prompting method, or "*Let's put speaker into interlocutor's shoes.*", surpasses the performance of Zero-shot CoT, especially in terms of emotion and intent accuracy, with an improvement of 21% and 7% respectively. The source code will be released after publication.

1 Introduction

Recent studies have witnessed the success of Chain-of-Thought (CoT) prompting [1, 2, 3] in achieving remarkable zero-/few-shot performance on complex reasoning tasks, including arithmetic, symbolic, and multi-modal [4], which benefited from providing step-by-step reasoning, called *rationale*, into Large Language Models (LLMs) [5, 6, 7, 8, 9]. In a zero-shot setting, a standard approach of Zero-shot CoT [10] 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 [11], and Theory-of-Mind (ToM) [12]. 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 [13], 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 [14, 15], it plays a crucial role in empathetic communication, referred to as Active Empathetic Listening (AEL) [16]. 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* [17, 18, 19], which is the act of perceiving and understanding another person's situation by putting ourselves in the other's shoes.

In this study, we explore the potential of Zero-shot CoT to enhance the empathetic reasoning abilities of LLMs. Through our experiments, we demonstrate that Zero-shot CoT is less effective in unlocking the 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.*") outperforms Zero-shot CoT, especially in terms of EMOACC and INTENTACC, achieving performance gains of 21% and 7% respectively.

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

2 Related Work

Chain-of-Thought Prompting. Recently, Chain-of-Thought (CoT) prompting [1, 2, 3] 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 [10], 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 [11], and Theory-of-Mind (ToM) [12]. 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 [20], many studies have proposed social dialogue generative agents, specifically to express empathy in social dialogues, by leveraging a mixture of experts [21], mimicking the interlocutor's emotions [22], commonsense knowledge [23], causality [24], and the Rational Speech Acts (RSA) framework [19]. Furthermore, a recent study [25] has demonstrated that GPT-3 [26], in a zero-/few-shot setting, achieved better performance than Blender 90M [27] 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 Method

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:

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

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.

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 [28].

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

Models	Prompting	Diversity		Empathy							
		Dist-1	Dist-2	IP	EX	ER	diff-IP	diff-EX	diff-ER	EMOACC	INTENTACC
text-davinci-001	ICL	0.0762	0.3875	0.2640	0.4120	0.8367	0.8507	1.0347	0.7753	0.1633	0.2707
	Z-CoT	0.1338	0.5421	0.2622	0.5081	0.6402	0.8577	1.2073	0.8659	0.1596	0.2571
	Δ	0.0576	0.1546	-0.0018	0.0961	-0.1965	0.007	0.1726	0.0906	-0.0037	-0.0136
text-davinci-002	ICL	0.0745	0.3547	0.2760	0.2480	0.9807	0.8693	0.8453	0.9287	0.1513	0.2740
	Z-CoT	0.1049	0.4387	0.2268	0.5732	0.8911	0.8357	1.3214	1.0196	0.1509	0.2277
	Δ	0.0304	0.084	-0.0492	0.3252	-0.0896	-0.0336	0.4761	0.0909	-0.0004	-0.0463
text-davinci-003	ICL	0.0593	0.3286	0.128	0.976	1.106	0.7600	1.8053	0.9633	0.1800	0.2647
	Z-CoT	0.0783	0.3591	0.0888	0.9882	0.8952	0.7078	1.8959	0.8668	0.1666	0.2346
	Δ	0.019	0.0305	-0.0392	0.0122	-0.2108	-0.0522	0.0906	-0.0965	-0.0134	-0.0301

Table 1: **Zero-shot Results.** We evaluate the zero-shot performance of LLMs using different prompting methods (i.e., ICL and CoT) across various metrics. A red cell indicates a performance decrease when Zero-shot CoT is applied. Z-CoT and Δ represent Zero-shot CoT and the difference in performance between Z-CoT and ICL, respectively.

Template	Empathy							
	IP	EX	ER	diff-IP	diff-EX	diff-ER	EMOACC	INTENTACC
Let's think step by step.	0.2268	0.5732	0.8911	0.8357	1.3214	1.0196	0.1509	0.2277
Let's think step by step from `[I]`'s perspective.	0.2279	0.5512	0.8401	0.8587	1.2933	1.0088	0.1431	0.2235
Let's think step by step, putting ourselves in `[I]`'s shoes.	0.2399	0.5886	0.8718	0.8708	1.3173	1.0249	0.1448	0.2362
Let's put ourselves in `[I]`'s shoes.	0.2522	0.3679	0.9827	0.8946	1.019	1.0484	0.1537	0.2401
Let's put `[S]` in `[I]`'s shoes.	0.4045	0.2921	1.0609	1.1199	0.9213	1.1685	0.3652	0.2921

Table 2: **Effect of Prompt Template.** We present the zero-shot performance of `text-davinci-002` using varied prompt templates on Empathy metrics. A green cell represents the highest performance for each evaluation metric. `[S]` and `[I]` denote the speaker and interlocutor names, respectively. Each name is randomly sampled from the Top-10K common names of US SSN applicants from 1990 to 2021, as described in Section 3.

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.

4 Experiments

4.1 Experimental Setup

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 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.

Language Models. We evaluate a total of three models with different zero-shot prompting methods: variants of InstructGPT (text-davinci-001, davinci-002, and text-davinci-003) [2] and different prompting methods (ICL and Zero-shot CoT).

Implementation Details. We conduct all experiments on a single A 100 (40GB) GPU. 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`.

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

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!

ICL:

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

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.

Perspective-Taking:

[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 3: **Case Study.** We present results generated by `text-davinci-002` with the Zero-shot ICL, Zero-shot CoT, and Zero-shot Perspective-Taking prompt methods for a given dialogue.

Evaluation Metrics. A standing issue of the empathetic dialogue generation task is to measure the empathetic reasoning ability due to the subjective nature of the empathy concept. Thus, we assess the LLMs' empathetic reasoning skills using various metrics related to Diversity, and Empathy, followed by the prior work [25]. For Diversity, we use Dist-*n* [29, 30]. For Empathy, we measure various metrics: EPITOME, DIFF-EPITOME, EMOACC, and INTENTACC. EPITOME [31] measures IP, EX, and ER by leveraging fine-tuned RoBERTa [32] model, respectively. We describe the EPITOME in Appendix B. DIFF-EPITOME [25] measures the difference scores of IP, EX, ER between the human golden response and predicted response, an extended version of EPITOME. EMOACC measures emotion accuracy using a fine-tuned BERT-base model [33] on the EMPATHETICDIALOGUES dataset labeled with 32 emotion categories. INTENTACC measures the response intent accuracy using a fine-tuned BERT model on the EMPINTENT dataset [34].

4.2 Results

Zero-shot CoT vs. Zero-shot ICL. Table 1 shows the zero-shot performance of three LLMs measured regarding Diversity and Empathy depending on the prompting methods (i.e., ICL and Zero-shot CoT). Zero-shot CoT generally fails to enhance zero-shot performance of LLMs across most evaluation metrics, particularly in Empathy (refer to Δ in Table 1). Despite improving EX performance, Zero-shot CoT considerably reduces diff-EX performance, implying that rationale compels LLMs to generate responses regarding the emotional situation of the interlocutor excessively. However, regarding Diversity, Zero-shot CoT tends to yield more varied responses, benefiting from the rationale.

Effect of Prompt Template. To evaluate the empathetic reasoning ability of LLMs compared to Zero-shot CoT, we compare the zero-shot performance of various prompt templates as demonstrated in Table 2. Generally, prompts related to *perspective-taking* show improved performance across most metrics. Notably, a prompt providing explicit information about speaker and interlocutor names, such as "*Let's put `[S]` in `[I]`'s shoes.*", surpass Zero-shot CoT significantly by approximately 21% in EMOACC and about 7% in INTENTACC. This *perspective-taking* prompt also performs better than Zero-shot ICL. These findings suggest that understanding an interlocutor's perspective is crucial to both LLMs and humans in empathetic reasoning.

Case Study. Table 3 presents generated responses and rationales by `text-davinci-002` using different prompting methods, such as ICL, CoT, and Perspective-Taking, in a zero-shot setting. Compared to the golden response, the Perspective-Taking prompting method makes the LLM generate more human-like empathetic responses, such as the reaction "Eww" and understanding gross emotion, resulting in improved performance (reported in Table 2). However, Zero-shot CoT generates responses (i.e., "B.") that are unreasonable and closer to the answers found in a multiple-choice setting. Interestingly, Zero-shot CoT attempts to understand the interlocutor's emotional situation step-by-step in the generated rationale, even the generated response from stage 2 of the Zero-shot CoT framework. We present the prompt templates used in our experiments in Appendix A.

5 Discussions

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 [31, 19, 25] 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 [33]. These metrics can cause inaccurate performance and are insufficient to evaluate the diverse and high-quality responses generated by InstructGPT [5]. Recently proposed prompt-based evaluators [35, 36] might help but can also prefer the LLM-generated responses, reported in a prior work [36]. 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 [19], the RSA framework [37] has previously been used to show an increase in empathetic dialogue generation across various dialogue generative models, such as MIME [22], DodecaTransformer [38], and Blender [27]. 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.

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 EMOACC and INTENTACC, surpassing both Zero-shot ICL and 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.

References

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

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

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

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

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

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

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

[8] OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023.

[9] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

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

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

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

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

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

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

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

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

[18] Perrine Ruby and Jean Decety. 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, 2004.

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

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

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

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

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

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

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

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

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

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

[29] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[30] Seungju Han, Beomsu Kim, and Buru Chang. Measuring and improving semantic diversity of dialogue generation. *arXiv preprint arXiv:2210.05725*, 2022.

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

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

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

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

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

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

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

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

A Prompt Template

As shown in Figure 1, we present prompt templates of Zero-shot ICL, Zero-shot CoT, and Zero-shot Perspective-Taking used in our experiments.

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:
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 put Demetrius in Jaylen's shoes.

Figure 1: **Prompt Templates.** A prompt template for Zero-shot ICL (**top**). A prompt template for Zero-shot CoT (**middel**). A prompt template for Zero-shot Perspective-Taking (**bottom**).

B EPITOME-based Automatic Evaluation

EPITOME, introduced by [31], 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.

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
ICL: **[Response]**: Yes, I can imagine that.
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.
Perspective-Taking: **[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 4: **Case Study 2.** We present results generated by `text-davinci-002` with the Zero-shot ICL, Zero-shot CoT, and Zero-shot Perspective-Taking prompt methods for a given dialogue.

In a recent study [19], each mechanism was used as an automatic metric to measure the empathy of generated responses using a fine-tuned RoBERTa [32] 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 4.