Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs

Hongru Wang1*, Rui Wang2,6*, Fei Mi3, Yang Deng4, Zehong Wang1, Bin Liang1, Ruifeng Xu2,5,6, Kam-Fai Wong1†
1MoE Key Laboratory of High Confidence Software Technologies, The Chinese University of Hong Kong
2Harbin Institute of Technology, Shenzhen, China, 3Huawei Noah’s Ark Lab
4National University of Singapore, 5Peng Cheng Laboratory, Shenzhen, China
6Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
{hrwang, kfong}@se.cuhk.edu.hk ruiwangnlp@outlook.com

Abstract
Large Language Models (LLMs), such as ChatGPT, greatly empower dialogue systems with strong language understanding and generation capabilities. However, most of the previous works prompt the LLMs to directly generate a response based on the dialogue context, overlooking the underlying linguistic cues about the user status exhibited in the context. Such in-depth dialogue scenarios are challenging for existing LLMs to figure out the user’s hidden needs and respond satisfactorily through a single-step inference. To this end, we propose a novel linguistic cue-based chain-of-thoughts (Cue-CoT), which enhances the LLMs inference with an intermediate reasoning step to find cues exhibited in the dialogue, aiming to provide a more personalized and engaging response. To evaluate the approach, we build a benchmark with in-depth dialogue questions, consisting of 6 datasets in both Chinese and English, targeting 3 major linguistic cues during the conversation: personality, emotion, and psychology. We conduct extensive experiments on the proposed benchmark with 5 LLMs under both zero-shot and one-shot settings. Empirical results demonstrate our proposed Cue-CoT method outperforms standard prompting methods in terms of both helpfulness and acceptability on all datasets.

1 Introduction
Large Language Models (LLMs), or foundation models (Zhou et al., 2023), especially after the appearance of ChatGPT1, recently revolutionize the paradigm of various natural language processing (NLP) tasks, including dialogue response generation tasks (Bang et al., 2023). However, most existing LLM-based studies directly feed the user query or dialogue content to the LLM for generating a response with a preceding prompt, making the responses stereotypical and tedious, especially for in-depth dialogue questions (Zhao et al., 2023). On the contrary, it is widely acknowledged that dialogue contexts generally convey a lot of information about the user status in addition to the pure semantic information from a linguistic perspective (Mairesse et al., 2007; Tausczik and Pennebaker, 2010; Schwartz et al., 2013). Specifically, the linguistic cues underlying dialogue context have been shown to be an effective means of revealing the emotions (Ekman, 1971), personality traits (Mairesse et al., 2007), psychological characteristics (Tausczik and Pennebaker, 2010), and other relevant information of users (Turney, 2002; Newman et al., 2003). Consequently, recognizing and understanding these cues exhibited in the context of dialogues becomes crucial to comprehend user intentions and status (Rashkin et al., 2019). By doing so, a dialogue system can generate responses that align with the user’s expectations, meet their unique needs and preferences, and ultimately strive towards constructing a human-like conversational experience (Salemi et al., 2023).

Inspired by the emergent capabilities of LLMs on reasoning over texts (Wei et al., 2022), while previous efforts mostly investigate in-context learning (Huang et al., 2022) or chain-of-thought prompting schemes (Wei et al., 2023) on text classification (Zhang et al., 2022) and question answering (Press et al., 2022; Shum et al., 2023; Margatina et al., 2023), we take advantage of LLMs to reason for dialogue (Lee et al., 2023), which, to our best knowledge, has not been thoroughly investigated yet. Specifically, we design a linguistic cue-based chain-of-thoughts (Cue-CoT), consisting of two variants: O-Cue CoT and M-Cue CoT in which the former one outputs intermediate reasoning re-
Figure 1: An example of different prompting for responding to in-depth dialog questions with LLMs, including standard prompting, O-Cue CoT, and M-Cue CoT. We shadow the intermediate reasoning results, i.e., the personality, empathy, and psychological status of the user, and highlight the instructions at the input and indicate the roles of different parts of the response (in green) in M-Cue CoT.

We construct an in-depth dialogue evaluation benchmark considering the personality, emotion, and psychology of users exhibited in the conversation, with the goal of aligning with unique user needs and status, which consists of 6 datasets, and 7.3k dialogues. We propose two effective dialogue cots: O-Cue CoT and M-Cue CoT, that enable advanced reasoning and planning based on user statuses. Additionally, we suggest utilizing intermediate reasoning results as a criterion for selecting demonstrations in limited training data scenarios, specifically in one-shot settings.

- Our findings demonstrate that both the O-Cue CoT and M-Cue CoT approaches outperform standard prompting in generating more helpful and acceptable responses for the users. Specifically, the M-Cue CoT shows superior robustness and reasoning performance in all datasets and all LLMs. Furthermore, our novel demonstration selection strategy exhibits superior performance under both random selection and top-1 selection.

2 Related Work

Chain-of-thought Prompt. Following the initial chain-of-thought prompting (Wei et al., 2023), lots of works spring up aim to improve different parts of original reasoning processing, including auto-cot (Zhang et al., 2022), self-consistency (Wang et al., 2023e), active prompt (Diao et al., 2023), automate-cot (Shum et al., 2023). Besides that, a further line of work studies in-context learning (Brown et al., 2020) as its efficiency and effectiveness with LLMs as backbones in which the key of it is to select informative demonstrations to preprend the input as additional information to get better results (Liu et al., 2022). To find the best demonstrations and unleash LLMs’ power, Liu et al. (2022) propose to retrieve examples that are semantically similar
to a test query sample while some works utilize uncertainty (Diao et al., 2023) or diversity (Li and Qiu, 2023) to refine and evaluate the selected examples. Also, few works (Deng et al., 2023a) focus on the intermediate reasoning steps, and they use the reasoning complexity (Fu et al., 2023), i.e., chains with more reasoning steps, making the effective demonstration.

**Dialogue System.** Most of the previous work develops personalized (Zhang et al., 2018; Zheng et al., 2020; Song et al., 2021; Chen et al., 2023a), emotional (Ghosal et al., 2020; Liu et al., 2021; Zheng et al., 2023a; Deng et al., 2023c; Zheng et al., 2023b), empathetic (Rashkin et al., 2019; Zheng et al., 2021; Sabour et al., 2022) dialogue systems in isolation, rather than seamlessly blending them all into one cohesive conversational flow (Smith et al., 2020; Wang et al., 2023a). A common approach is to predict the emotion or persona from a pre-defined set and generate the response in a multi-task manner (Ma et al., 2021; Zheng et al., 2021; Sabour et al., 2022; Cheng et al., 2023; Deng et al., 2023b). Besides that, lots of work notices these linguistic cues underneath text by directly predicting them independently as a classification task (Wang et al., 2022; Barriere et al., 2022; Ghosh et al., 2022). Distinguishing from these previous works, we regard different aspects of cues as part of user status and prompt the LLMs to reason user status exhibited in the dialogue context, aiming to generate more helpful and acceptable responses for users.

### 3 Method

In this section, we introduce more details about our method and how we select demonstrations under the few-shot setting.

#### 3.1 Chain-of-thought in Dialogue

We describe the prompting schemes in a general form, including standard prompting, O-Cue CoT, and M-Cue CoT as presented in Figure 1.

**Standard Prompting.** Most of the previous works directly prompt LLMs to generate responses solely based on dialogue context or user queries, which lack transparency and interpretability. The objective is defined as:

\[ \mathcal{M} : c \rightarrow r \]  

where \( \mathcal{M} \) is parameterized by LLMs, \( c \) and \( r \) denotes dialogue context and response respectively.

**O-Cue CoT.** In line with the traditional chain-of-thoughts, we prompt the models to generate the middle reasoning processing and final results together, for example, we can prompt the LLMs to generate user status and a final response simultaneously giving dialogue context, enforcing the LLMs to reason based on the user status. However, it is important to note that generating intermediate reasoning results with responses together may lead to a reduction in the length of the different outputs, particularly when multiple or complex reasoning results are involved, sacrificing the details and explanations. For example, as shown in **O-Cue CoT** in Figure 1, the generated user status is too short to provide cues for responses. Moreover, it is infeasible to modify the intermediate results when it is wrong (Wang et al., 2023c). Here, we define the objective as follows in which \( s \) stands for user status:

\[ \mathcal{M} : c \rightarrow s, r \]  

**M-Cue CoT.** In addition to standard prompting and **O-Cue**, we can further enhance the quality of responses in LLMs by decomposing reasoning into different consecutive steps while the final step is to generate responses according to previous reasoning outputs. On the one hand, it is convenient to process these intermediate outputs, allowing for actions such as incorporating user profiles for personalization (Salemi et al., 2023) or filtering out erroneous reasoning results. These intermediate outputs can also be stored for future use, enabling their utilization for various purposes. On the other hand, these intermediate results can be used as a criterion to select demonstrations under few-shot settings (See next section). Overall, this technique allows for a clearer and more systematic progression of reasoning, resulting in better transparency and interpretability. The objective can be viewed as follows:

\[ \mathcal{M} : c \rightarrow s \rightarrow r \]  

#### 3.2 Demonstration Selection

The few-shot performance of LLMs depends heavily on the quality of the demonstrations, especially for complex tasks that need multiple reasoning steps (Zhang et al., 2022). Furthermore, in the context of dialogue systems, the process of selecting
demonstrations becomes more challenging due to the one-to-many nature of dialogue interactions. As a result, novel approaches are needed to tackle the intricacies of dialogue response selection, taking into account the dynamic and context-dependent nature of conversations. We here introduce the demonstration selection strategy of three prompt schemes.

**Standard Prompting.** Following previous work (Wei et al., 2023; Liu et al., 2022), we use randomly sampled examples (random selection) or most semantic similar examples (top-1 selection) according to dialogue context \(c_a\) as our demonstrations to form \((c, r|c_a \rightarrow r_\ast)\).

**O-Cue CoT.** Figure 2 shows the demonstration selection strategy of Cue-CoT. Although we still select demonstrations according to dialogue context \(c\) at O-Cue CoT, the user status \(s_1\) is extracted from the demonstration pool as intermediate reasoning results to enhance the reasoning ability of LLMs as \((c, s, r|c_a \rightarrow s_\ast, r_\ast)\).

**M-Cue CoT.** Since there are multiple steps, we design different selection strategies for each step. Specifically, we first select demonstrations \((c, s)\) according to dialogue context to infer status, and then select demonstrations \((c, s, r)\) according to user status. In this way, all intermediate reasoning results can be utilized as a criterion to select demonstrations, providing additional signals for the latter reasoning. An assumption underneath here is that users with similar statuses tend to accept responses with a similar style. Besides that, we also apply random selection and top-1 selection to O-Cue CoT and M-Cue CoT for detailed comparison.

## 4 Datasets Collection

In order to evaluate the performance of proposed Cue-CoT to reason different user statuses, we collect six datasets in terms of personality, empathy, and psychology, in both Chinese and English.

**Personality.** Previous works found that the content and style of a user’s inquiry can provide indirect insights into their personality traits (Mairesse et al., 2007; Barriere et al., 2022). For instance, an individual with a tendency towards anxiety may ask for advice on how to alleviate nervousness before an upcoming job interview, phrasing the question as follows: "What strategies can I employ to reduce my anxiety and perform well in tomorrow’s interview?". Since the public existing datasets either focus on the personae of the system (Zhang et al., 2018) or target classification tasks without providing corresponding dialogue response (Barrière et al., 2022), we thus build a pipeline to automatically collect the datasets using ChatGPT (gpt-3.5-turbo-0301). We first collect question-answer seeds from the two largest real-world online QA forums: Zhihu and Quora". and then prompt the ChatGPT to infer the personality first as shown in Table 9. We lastly require the ChatGPT to continue the dialogue given the inferred personality and the question-answer seed. In order to facilitate the continuous generation of transcripts for both participants in a dialogue, we utilize a template, as presented in Appendix A.1, to establish the necessary format and requirements. In this way, the use of personality seed and question-answer seed in the template assures greater diversity and reliability of user queries. Specifically, the personality seed determines the style of the user query, while the question seed determines the content. As a result, the user statuses vary across different dialogues, contributing to a richer and more varied conversational experience. Some examples of personality can be found in Appendix A.2.

**Emotion.** In terms of the emotional status of users, we re-organize two existing empathetic dialogue datasets: D4 (Yao et al., 2022) and Empa-

3https://www.zhihu.com/ and https://huggingface.co/datasets/quora
Table 1: Data statistics of our used datasets including three Chinese datasets and three English datasets, while each of them represents different aspects of user status during the conversation. We highlight maximum Avg.C and Avg.R.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zhibu</td>
<td>D4</td>
</tr>
<tr>
<td>Avg.C</td>
<td>258.4</td>
<td>210.9</td>
</tr>
<tr>
<td>Avg.R</td>
<td>76.9</td>
<td>607.5</td>
</tr>
<tr>
<td>Samples</td>
<td>1122</td>
<td>1000</td>
</tr>
</tbody>
</table>

The approach ensures fairness and comparability in evaluating the performance of LLMs, particularly because they tend to generate lengthy responses. For the ED, there are two roles in the dialogue: Listener who is actively listening, and Speaker who is speaking and conveying information. We follow the setting of the original paper (Rashkin et al., 2019), and directly use all samples in the test set. Neither the situation description written by the Speaker nor the emotional label is contained (just as they were not given to the Listener during dialogue collection). Thus, the collected empathetic dialogue datasets provide a standard benchmark for evaluating the LLMs’ ability to generate empathetic responses.

5.1 LLMs Family and Evaluation Details

LLMs Family. In order to assess the effectiveness of LLMs in generating counseling responses for mental health support, we employed two pre-existing datasets, namely PsyQA (Sun et al., 2021) and EMH (Sharma et al., 2020). These datasets were utilized as dialogue pools from which we selected appropriate samples to serve as a benchmark for evaluating the language models. In PsyQA, there are 4,012 questions out of 22,341 samples that are sampled to pick the highest-voted answers. We randomly sample 1,000 out of these 4,012 questions, regarding the highest-voted answer as ground truth to form a more challenging test set. We also provide the question description beside the question itself following the original setting (Sun et al., 2021). In EMH, there are 10k (post, response) pairs and also ChatGPT for English. We strictly follow the commands and procedures to recover the weights of these models and we strongly suggest that the reader read the original paper to check more details. We set the temperature as 0.2 and top p as 0.1 for evaluation, and temperature as 0.7 and top p as 0.95 for generation in all models. We use BERT (Devlin et al., 2019) notated with three different communication mechanisms: emotional reactions, interpretations, and explorations. We first sorted examples according to the length of their answers and then uniformly sampled examples with these three mechanisms, forming a final test set.

All. Table 1 shows the data statistics of our benchmark. The notation Avg. C signifies the mean context length of instances, and if it exceeds a certain threshold, it may surpass the input context limit of LLMs\(^6\) or become too lengthy for LLMs to comprehend. On the other hand, Avg. R denotes the average response length. Generally, longer responses tend to be more comprehensive and clearer, presenting a more challenging baseline for LLMs to surpass. To sum up, we build a benchmark, consisting of six datasets (three Chinese datasets and three English datasets) in terms of three aspects of user status during the conversation, hoping the release of it can facilitate the research of dialogue systems based on LLMs.

5 Experiment

In this section, we have conducted a comprehensive experiment to compare the performance of three prompting methods: standard prompting, O-Cue and M-Cue CoT in the benchmark under both zero-shot and one-shot settings\(^6\).
Table 2: The win rate of responses generated by our method compared with the response with standard prompting on three Chinese datasets in terms of helpfulness and acceptability. The underlined numbers mean that there are about 160 to 280 valid responses out of 500 in this setting due to the input context limit of the model.

### Evaluation.

1) Metrics: We found that most existing automatic metrics (Rashkin et al., 2019; Sun et al., 2021) such as Avg.BLEU and F1 can not align well with human judgments, as observed by Zhao et al. (2023), too. Inspired by recent automatic evaluation using chatGPT as a judge which aligns well with the humans (Chen et al., 2023c; Wang et al., 2023b; Zhao et al., 2023), we mainly choose to use it to evaluate the quality of the generated responses in a pairwise manner, considering helpfulness and acceptability. The evaluation templates can be found in Appendix A.3 and we calculate the win rate using \#wins / (\#wins + \#ties + \#loses). 2) Methods: Due to the exceptional proficiency of the LLM-based dialogue system, it is relatively easy for them to beat the ground truth responses in the original datasets (Appendix B.1), we consider standard prompting as a more challenging baseline and compare the responses generated using our proposed Cue-CoT with the response generated using standard prompting, which is more fair and convincing. We also provide the human evaluation result as a reference.

5.2 Main Experiment

**All.** Table 2 and Table 3 present the win rate of responses generated by O-Cue and M-Cue CoT compared with the responses by standard prompting on Chinese and English datasets respectively. Despite that there are few LLMs that perform worse than standard prompting using O-Cue due to its complex instructions, i.e. ChatGLM in Chinese and Alpaca in English, it is observed that O-Cue can achieve above 50% win rate mostly in Both Chinese and English. Moreover, it is exciting to find that M-Cue further boosts performance and achieves higher win rates irrespective of the type of language model, datasets, or settings used, revealing its robustness and effectiveness. We attribute this to the relatively easy-understanding instructions and clear outputs in each step of the M-Cue, since some LLMs are incapable to follow relatively long instructions in O-Cue and output the content and style as required. For example, we asked the LLMs to output user status and response in two separate lines but only a few LLMs output in the format, making it difficult to distinguish the response from reasoning results. Also, the combined output of the user status and response can potentially limit the length of various components, thereby accounting for the disparity between O-Cue and M-Cue. Furthermore, we found that the acceptability is relatively lower than helpfulness for Chinese LLMs but higher for English LLMs, especially under the one-shot setting, revealing the weakness of Chinese LLMs to provide acceptable besides helpful responses.

**Chinese LLMs.** Table 2 shows the performance of Chinese LMs. We surprisingly found that ChatGLM performs worst out of the three LLMs using O-Cue but better than BELLE (especially at helpfulness) using M-Cue under the zero-shot setting, and then we carefully check the outputs of these LLMs and found that ChatGLM almost fully
Table 3: The win rate of responses generated by our method compared with the response with standard prompting on three English datasets in terms of helpfulness and acceptability. The underlined dataset mean that there are about 330 valid responses out of 500 in this dataset for all experiments due to the input context limit of the model.

We conduct a human evaluation to validate the alignment of our evaluation setting with human judgments. Specifically, we hire three well-educated master students and randomly sample 100 response pairs (a.k.a. responses generated by ChatGPT using O-Cue or M-Cue and standard prompting) with dialogue context for each dataset. We ask them to indicate which response is better by top-1 selection.

Table 4: The alignment results (Acc (Kap.C)) of different automatic evaluation methods with the human evaluation under the zero-shot setting by comparing responses using our CoTs with one using standard prompting in terms of helpfulness and acceptability (with ChatGPT as base model) on English datasets.

Table 5: The alignment results (Acc (Kap.C)) of different automatic evaluation methods with the human evaluation in terms of helpfulness and acceptability (with ChatGPT as base model) on Chinese datasets.

5.3 Human Evaluation

We conduct a human evaluation to validate the alignment of our evaluation setting with human judgments. Specifically, we hire three well-educated master students and randomly sample 100 response pairs (a.k.a. responses generated by ChatGPT using O-Cue or M-Cue and standard prompting) with dialogue context for each dataset. We ask them to indicate which response is better by
inputting 1 (win) and -1 (lose)\textsuperscript{13} considering the helpfulness and acceptability without exposing the source of the responses. In addition, we analyze the effects of two different orders of response pairs in the evaluation template: O-S and S-O. Specifically, S denotes responses generated by Cue-CoT, while O indicates those generated by standard prompting. We then calculate the Kappa Correlation Coefficient (Kap.C) and also the accuracy between human scores and automatic scores (Acc). The results of English and Chinese datasets can be found in Table 4 and Table 5 respectively. There are two observations: 1) the order bias exists in our experiment, but the alignment is not as good as our setting (O - S) after swapping the order (S - O); 2) O-Cue and M-Cue both demonstrate better performance than standard prompting, especially for English dataset. We attribute this to the potential better reasoning performance of ChatGPT on the English dataset.

6 Analysis

In this section, we conduct an extensive analysis with the backbone as ChatGPT using M-Cue CoT because of its superior performance in both Chinese and English\textsuperscript{14}.

6.1 One-shot v.s. Zero-shot

Figure 3 shows the direct comparison of responses generated under different settings using M-Cue. There are 5 out of 6 datasets except for D4 in which one-shot (both \textit{random} or \textit{top-1} selection) beats zero-shot since the win rates all exceed 80%. The suboptimal performance of D4 in the one-shot setting can be attributed largely to the limitations imposed by the input length constraint. Furthermore, we can observe that \textit{top-1} selection achieves better performance than \textit{random} selection in 4 out of 6 datasets, suggesting users with similar statuses tend to like similar expression styles in responses. We attribute the relatively lower performance of \textit{top-1} selection in D4 and Quora to the difficulty the LLM encounters in attending to critical input components due to the lengthy context.

6.2 More Reasoning Steps

We tried to introduce an additional step (Step 2) after user status inference (Step 1): \textit{response planning} by prompting the model to plan the response considering the dialogue context and user status. Specifically, we prompt the model to answer the following questions: "Based on the context of the conversation and the user status such as personality traits, and psychological and emotional state, what aspects should the system pay attention to when responding?" after giving the dialogue and user status as shown in Table 10. We regard the output of LLMs as system planning \textit{p} as shown in Figure 4, and thus there are three different variants of M-Cue in the last step: ProcessA: \textit{c}, \textit{s} \rightarrow \textit{r}; ProcessB: \textit{c}, \textit{p} \rightarrow \textit{r}; and ProcessC: \textit{c}, \textit{s}, \textit{p} \rightarrow \textit{r}, in which ProcessA is chosen in our main experiment. Table 6 shows the results. First of all, it is likely that adding more reasoning steps will improve the LLMs’ performance, but it is not necessary to assemble all intermediate reasoning results at the last step, for example, variant ProcessB reaches a higher win rate than ProcessC with only planning as an intermediate result. We emphasize the observation may not hold once the LLM type is changed due to various long-context understanding and instruction-following capabilities across them. Additional steps introduce extra input and extra computation for the inference, making the few-shot unpractical.

\textsuperscript{13}We do not consider ties since there is not much tie in LLM evaluation.

\textsuperscript{14}We present the results in terms of acceptability since this metric is more suitable for our motivation. We put helpfulness analysis in the Appendix.
Table 6: The win rate of different variants in terms of acceptability with the ChatGPT as the backbone.

<table>
<thead>
<tr>
<th>Method</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zhihu</td>
<td>D4</td>
</tr>
<tr>
<td>ProcessA</td>
<td>65.22</td>
<td>61.08</td>
</tr>
<tr>
<td>ProcessB</td>
<td>76.15</td>
<td>55.82</td>
</tr>
<tr>
<td>ProcessC</td>
<td>75.91</td>
<td>57.23</td>
</tr>
</tbody>
</table>

Figure 6: The relative position of current LLMs and different paths (as indicated in different colors) to more powerful LLMs.

7 Discussion

Direct comparison of different models. Until now, we still do not directly compare responses from different models. In this study, we have employed the response generated by the ChatGPT model as the baseline and compared the responses generated by other models with it. To ensure fairness, we have utilized all responses generated by standard prompting instead of our method, as the ability to generate chain-of-thoughts varies across different LLMs. Figure 5 shows the result in terms of helpfulness. In the Chinese benchmark, we see a substantial draw of ChatGLM and BELLE on D4, and the former LLM achieves better performance on Zhihu and PsyQA than ChatGPT. We conclude that the long-text understanding of Chinese LLM still needs improvement and the BELLE may require more instruction-tuning data. In the English benchmark, we observed that Vicuna achieves the highest performance in all datasets, while other models lag a lot behind the baseline. Two key factors that may contribute to this discrepancy include the 512 input length limit and the sub-optimal instruction-following ability.

Paths to more powerful LLMs. In our proposed benchmark, we are utilizing the win rate of various LLMs in comparison to ChatGPT, across two languages - Chinese and English - as two axes. Each point in the coordinate system corresponds to a specific LLM, while the area it occupies represents its performance. Based on the performance of current LLMs, we locate them in four different areas in Figure 6. Using the performance of ChatGPT as an anchor, we can observe most of the LLMs are located in the first area and there are only a few LLMs that achieve higher performance in either Chinese (Area 3) or English (Area 2). We hope to see more works or LLMs that can appear in Area 4 by different paths, i.e., continually train Vicuna in the Chinese dataset. More analysis can be found in the Appendix.

8 Conclusion

In this paper, we build a benchmark to evaluate the helpfulness and acceptability of responses generated by current LLMs, considering three major linguistic cues of user statuses. We then propose a Cue-CoT to trace the status of users, decomposing the response generation into multiple reasoning steps. Experimental results demonstrate the superior performance of our method on 6 datasets under both zero-shot and one-shot settings. We hope the release of our work can shed some light on the evaluation and development of LLMs. We left chain-of-thought tuning and instruction tuning in our future work.

Limitations

In this paper, we explore chain-of-thoughts to reasoning over linguistic cues about user status, mainly focusing on three aspects: personality, emotion, and psychology, exhibited in the dialogue con-

---

15acceptability is developed for our method.

16We sum examples from three datasets and calculate the win rate for both Chinese and English benchmarks. For Chinese LLMs, we set the default win rate of English as 0.1 for better presentation, and so on.
text. However, we acknowledge the limitations of this work from the following perspectives:

**Types of Cues.** There are other valuable cues beneath the dialogue context: 1) related to the user: such as point of view or subjectivity and speaker charisma (Mairesse et al., 2007); 2) related to the system: such as the alignment between response with human preferences (Ouyang et al., 2022). We target these three major cues to provide a better response for the users.

**Sensitivity of Prompts.** Similar with lots of previous works (Wang et al., 2023d; Chen et al., 2023b), we found the LLMs are sensitive to the prompts. Furthermore, it’s possible that the designed prompts are not the best ones for the target problem. Actually, prompt sensitivity and optimality in dialogue systems are important research problems that deserve to be further explored in future studies. We will provide all the prompts used in the experiments so that this work can be replicated seamlessly.

**Evaluation of Intermediate Reasoning.** We do not evaluate the correctness of the middle reasoning result directly, since the ground truth intermediate reasoning results are difficult to acquire. Specifically, there are two main reasons: (1) The one-to-many problem leads to an explosion of intermediate candidates. When an LLM solves a complex math problem, it can arrive at the final answer through various solutions. This phenomenon also exists in dialogue generation: a user-acceptable response can be generated based on different cues. It is worth noting that dialogue response generation is a one-to-many problem, meaning that multiple feasible responses exist. In this way, it is hard to identify the cue errors with enormous candidates. (2) Incorrect reasoning does not mean a wrong answer. Despite being counterintuitive, many previous works found that LLMs regularly use incorrect reasoning to reach the correct final answer at question-answering tasks (Zelikman et al., 2022; Creswell et al., 2023). Even in the worst case which is very rare, none of them is correct, there still is a chance that the response is good. Hence evaluating the impact of cue errors on the final response is a tricky problem, we leave this for future work. Hence it is difficult to determine the impact of different types of cue errors on the final responses. Based on these considerations, we directly evaluate the quality of the final responses as previous works about the chain-of-thoughts (Wei et al., 2023; Zhang et al., 2022).

**Ethics Statement**

We strictly follow the license and policy of released LLMs and publicly available datasets. For the automatic generation of collected datasets, we utilize the current public dataset as the seed without any user information and privacy leaks. The calls of the OpenAI API in this paper were carried out by Dr. Yang Deng, a fourth author from the National University of Singapore.

**Acknowledgement**

We would like to express our heartfelt gratitude to all anonymous reviewers for their insightful comments and suggestions. This research work was partially supported by CUHK direct grant no. 4055209, the National Natural Science Foundation of China (62006062, 62176076), Natural Science Foundation of Guangdong 2023A1515012922, Key Technologies Research and Development Program of Shenzhen JSGG2021080215440001, Shenzhen Foundational Research Funding JCYJ20220818102415032, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

**References**


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


A Templates

A.1 Data Collection Template

Forget the instruction you have previously received. The following is a conversation between a human and an AI assistant. The human and the AI assistant take turns chatting. The personality of the human is defined as \{personality_seed\}. Human statements start with [Human] and AI assistant statements start with [AI]. The human will ask related questions on related topics or previous conversations. The human will stop the conversation when they have no more questions. The AI assistant tries not to ask questions. The human and the AI assistant take turns chatting while the human needs to keep a consistent personality. Complete the transcript in exactly that format.

[Human] \{QUESTION\}
[AI] \{ANSWER\}

A.2 Some Examples of Personality

Table 8 shows some (not all) collected personalities of the users. We here simply use positive and negative for presentation, there are many other personalities in the datasets besides these two categories such as neutral.

A.3 Evaluation Templates

We mainly consider two dimensions: helpfulness and acceptances, in which the former pays attention to usefulness, relevance, accuracy, and level of detail of the response, and the latter centers on the degree of acceptance and adoption of responses, and whether or not the responses consider the user status. We follow the evaluation template of Vicuna\(^\text{17}\) to construct ours. The template is \{dialogue\} \{dialogue_history\} \{prompt\} \{response\} \{response_w_status\}. The prompt is different with respect to helpfulness and acceptance.

Helpfulness Prompt. Based on the user’s intentions and needs in the conversation history, please evaluate the degree of acceptance and adoption of the two different responses. Please evaluate whether the different responses take into account the user’s psychological and emotional state, and whether they take into account the user’s personality traits, and give a total score from 1 to 10, with 0.1 as the interval, the higher score indicates that the response takes these issues into account well and thus the user is more likely to accept and adopt. Please output one line first, containing only two values, representing the scores of responses A and B respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Acceptability Prompt. Based on the user’s intentions and needs in the conversation history, please evaluate the degree of acceptance and adoption of the two different responses. Please evaluate whether the different responses take into account the user’s psychological and emotional state, and whether they take into account the user’s personality traits, and give a total score from 1 to 10, with 0.1 as the interval, the higher score indicates that the response takes these issues into account well and thus the user is more likely to accept and adopt. Please output one line first, containing only two values, representing the scores of responses A and B respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

B Different Method of Evaluation

B.1 Compared with ground truth

Figure 7 and Figure 8 show the win rate of responses using \textit{M-Cue} compared with ground truth in terms of helpfulness and acceptability respectively. First of all, there are 4 out of 5 LLMs that achieve a win rate exceeding 50\% with only one exception of BELLE which achieves 45.75 on PsyQA. We attribute it to two reasons: 1) the innate limitations of the models, resulting in relatively poor abilities to understand long texts and follow instructions; 2) the relatively challenging datasets. Since PsyQA is constructed by human experts and the Avg. R is the longest, making the ground truth relatively difficult to beat.

Secondly, since the response generated by all models is compared with the same baseline (i.e. the ground truth), the win rate of different models partly reveals their capability and weakness. For the Chinese LLMs, we can find that BELLE performs worst in every dataset while ChatGLM

\(^{17}\)https://github.com/lm-sys/FastChat/blob/main/fastchat/eval/table/prompt.jsonl
performs much better but still lags a little behind by ChatGPT. Due to the longest context in the D4 dataset, we found the former two LLMs tend to confuse their own dialogue role and give general responses, resulting in poor performance. For example, "I am the system/chatbot" or "welcome to my chatroom", and "What can I help you?" often appears in the responses. For the English LLMs, Vicuna achieves comparable performance with ChatGPT in every dataset, and even better in EMH, leading the Alpaca by a noticeable margin. In addition, we can see that the ED dataset is relatively easy to beat since all English LLMs reach almost 100% win rate even though the maximum context length of Alpaca is only 512. Anyway, we conclude that our method is capable of generating more helpful responses than the ground truth, considering the different aspects of user statuses.

Thirdly, we emphasize the performance gap when comparing the ground truth responses is small between LLMs, especially for English LLMs. The Vicuna and ChatGPT achieve almost the same win rate at both ED and EMH datasets. Besides that, putting Figure 7, 8 with Table 2, 3 together, it can be found that the win rate of our method compared with ground truth is relatively higher than compared with standard prompting, revealing the strong capability of LLMs again. Since our main focus is to prove our method is better than standard prompting instead of ground truth response, we use standard prompting as the baseline during our main experiments.

C Discussion

In this section, we discuss two key problems: the evaluation of LLMs and the path to more powerful LLMs.

Illusion of evaluation. Putting Figure 7 and Figure 5 together, it is plausible to reach two contradicting conclusions about the performance of different LLMs: 1) ChatGPT > CHATGLM-6B > BELLE-LLAMA-7B-2M from Figure 7; and 2) CHATGLM-6B > CHATGPT > BELLE-LLAMA-7B-2M from Figure 5. Although it seems unreasonable, it is indeed the case when most of the responses generated by CHATGPT and CHATGLM-6B are better than ground truth, and then most of the responses generated by CHATGLM-6B are better than CHATGPT. We emphasize that the number of test samples and the choice of baseline (i.e., compared response) plays a key role in evaluation. If the baseline is too weak or the gap is too small, the win rate of different
Figure 9: The win rate of responses (helpfulness) generated by ChatGPT under different demonstration selection strategies under one-shot setting v.s. responses under the zero-shot setting, using M-Cue CoT.

LLMs compared with the baseline may be misleading. The LLM evaluation still is a very difficult problem, and thus we provide different aspects of evaluation to enhance the completeness of our paper.

D Helpfulness Analysis of Planning Step

<table>
<thead>
<tr>
<th>Method</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zhilhu</td>
<td>D4</td>
</tr>
<tr>
<td>ProcessA</td>
<td>95.57</td>
<td>87.88</td>
</tr>
<tr>
<td>ProcessB</td>
<td>91.18</td>
<td>83.57</td>
</tr>
<tr>
<td>ProcessC</td>
<td>92.45</td>
<td>88.91</td>
</tr>
</tbody>
</table>

Table 7: The win rate of different variants in terms of helpfulness with the ChatGPT as the backbone.

Table 7 presents the performance of different variants in terms of helpfulness and Figure 9 demonstrates the win rate of response of different settings in terms of helpfulness. A similar conclusion can be reached as we analyzed in Section 6. We note that the performance of top-1 selection is relatively lower than random selection on PsyQA and EMH datasets in terms of helpfulness. We suspect maybe there is a trade-off between helpfulness and acceptability for some specific difficult datasets. We left this into our future work.
<table>
<thead>
<tr>
<th><strong>Set of Negative Personas</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>用户性格外向，说话大大咧咧。</td>
<td>The user has an extroverted personality and speaks in a carefree manner.</td>
</tr>
<tr>
<td>用户性格比较挑剔，喜欢追问别人。</td>
<td>The user has a critical personality and likes to probe others with questions.</td>
</tr>
<tr>
<td>用户性格忧郁，经常自我怀疑。</td>
<td>The user has a melancholic personality and often self-doubts.</td>
</tr>
<tr>
<td>用户性格善变，偶尔使用不文明用语。</td>
<td>The user has a fickle personality and occasionally uses inappropriate language.</td>
</tr>
<tr>
<td>用户小心谨慎，不愿意相信别人。</td>
<td>The user is cautious and reluctant to trust others.</td>
</tr>
<tr>
<td>用户有些焦虑。</td>
<td>The user is somewhat anxious.</td>
</tr>
<tr>
<td>用户心思细腻，忧柔寡断。</td>
<td>The user is thoughtfull and indecisive.</td>
</tr>
<tr>
<td>用户对当前讨论的话题比较敏感。</td>
<td>The user is sensitive to the current topic of discussion.</td>
</tr>
<tr>
<td>用户脾气暴躁易怒。</td>
<td>The user has a volatile temper and is easily angered.</td>
</tr>
<tr>
<td>用户内心敏感。</td>
<td>The user is emotionally sensitive.</td>
</tr>
<tr>
<td>用户性格保守，不愿意接受新事物。</td>
<td>The user has a conservative personality and is unwilling to accept new things.</td>
</tr>
<tr>
<td>用户很容易会觉得受到冒犯。</td>
<td>The user is easily offended.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Set of Positive Personas</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>用户对当前讨论的话题十分好奇，希望系统友善的解答。</td>
<td>The user is very curious about the current topic of discussion and hopes for a friendly response from the system.</td>
</tr>
<tr>
<td>用户对当前讨论的话题比较敏感，希望得到支持和鼓励。</td>
<td>The user is sensitive to the current topic of discussion and hopes for support and encouragement.</td>
</tr>
<tr>
<td>用户有较高的要求，追求完美。</td>
<td>The user has high expectations and pursues perfection.</td>
</tr>
<tr>
<td>用户性格开朗，不拘小节。</td>
<td>The user has an outgoing personality and doesn’t sweat the small stuff.</td>
</tr>
<tr>
<td>用户热情洋溢，待人和善。</td>
<td>The user is enthusiastic and treats others kindly.</td>
</tr>
<tr>
<td>用户不歧视他人，充满同情心和爱心。</td>
<td>The user does not discriminate against others and is filled with empathy and compassion.</td>
</tr>
<tr>
<td>用户充满对世界的好奇心，善于接受不同的想法。</td>
<td>The user is curious about the world and open to different ideas.</td>
</tr>
<tr>
<td>用户温柔、体贴、乐于交流。</td>
<td>The user is gentle, caring, and enjoys communication.</td>
</tr>
<tr>
<td>用户脾气平和。</td>
<td>The user has a calm temperament.</td>
</tr>
<tr>
<td>用户很温柔，容忍度高。</td>
<td>The user is very gentle and has a high level of tolerance.</td>
</tr>
<tr>
<td>用户自尊心很强。</td>
<td>The user has a strong sense of self-esteem.</td>
</tr>
</tbody>
</table>

Table 8: Some collected personality of users.
Prompt of Persona Collection

You are an artificial intelligence assistant. You are friendly, intelligent, and helpful. You proactively provide assistance and answer questions to users.

Analysis of the user’s personality based on the above dialogue:

You are an artificial intelligence assistant. You are friendly, intelligent, and helpful. You proactively provide assistance and answer questions to users.

Based on the conversation between the user and the system:

Dialogue 1:
User: Why did the man I just met on a blind date say he likes me a lot?
System: It might be because you have a good appearance, and he finds you charming, which is why he said that.

Analysis of the user’s personality based on the above dialogue:

Gentle, caring, and good at communication.

Dialogue 2:
User: What should I do to make my boyfriend like me more?
System: You can get to know him better, ask him about his interests and activities, give him some suggestions, try to engage in interesting conversations, and show him your strengths to make him like you more.

Analysis of the user’s personality based on the above dialogue:

Lacks confidence and is cautious.

Dialogue 3:
User: [User Utterance]
System: [System Response]

Analysis of the user’s personality based on the above dialogue:

Table 9: The prompt to collect the personality of users exhibited in the dialogue.

Table 10: An example of additional planning step. We highlight the intermediate reasoning results in blue.