# Large Language Models With Holistically Thought Could Be Better Doctors

**Anonymous ACL submission**

## Abstract

The medical conversational question answering (CQA) system aims at providing a series of professional medical services to improve the efficiency of medical care. Despite the success of large language models (LLMs) in complex reasoning tasks in various fields, such as mathematics, logic, and commonsense QA, they still need to improve with the increased complexity and specialization of the medical field. This is because medical CQA tasks require not only strong medical reasoning, but also the ability to think broadly and deeply. In this paper, to address these challenges in medical CQA tasks that need to be considered and understood in many aspects, we propose the Holistically Thought (**HoT**) method, which is designed to guide the LLMs to perform the diffused and focused thinking for generating high-quality medical responses. The proposed HoT method has been evaluated through automated and manual assessments in three different medical CQA datasets containing English and Chinese languages. The extensive experimental results show that our method can produce more correct, professional, and considerate answers than several state-of-the-art (SOTA) methods, manifesting its effectiveness.

## 1 Introduction

*Broaden your thinking to gain multiple perspectives, deepen your thinking to gain insight.*

*Wise Man*

Medical research with AI-based techniques is growing rapidly (Khan et al., 2017; Gao et al., 2018; Javed et al., 2020; Li et al., 2022c). The medical conversational question answering (CQA) task is designed to improve the efficiency of medical care by providing a range of professional medical services for patients (Xu et al., 2019). The development of medical CQA systems has become an important research topic, particularly amid the COVID-19 outbreak (Yang et al., 2020; Zhou et al., 2021). Such a system can improve the patient experience during the clinical treatment process by quickly responding to their needs and providing information on symptoms, diagnosis, prescription of medications, and treatment suggestions (Palanica et al., 2019; Chen et al., 2021; Li et al., 2022a).

Medical research with AI-based techniques is growing rapidly (Khan et al., 2017; Gao et al., 2018; Javed et al., 2020; Li et al., 2022c). The medical conversational question answering (CQA) task is designed to improve the efficiency of medical care by providing a range of professional medical services for patients (Xu et al., 2019). The development of medical CQA systems has become an important research topic, particularly amid the COVID-19 outbreak (Yang et al., 2020; Zhou et al., 2021). Such a system can improve the patient experience during the clinical treatment process by quickly responding to their needs and providing information on symptoms, diagnosis, prescription of medications, and treatment suggestions (Palanica et al., 2019; Chen et al., 2021; Li et al., 2022a).

Although conventional language models have been employed in medical question-answering tasks (Wu et al., 2019; Liang et al., 2021), their efficacy in the medical domain is limited due to training on insufficient medical datasets (Li et al., 2021). Recent advances in large language models (LLMs) have demonstrated improved performance through in-context learning (Dong et al., 2022) and few-shot samples (Brown et al., 2020a). Additionally, the chain-of-thought (CoT) prompting method (Wei et al., 2022) has been shown to facilitate complex reasoning in medical tasks by decomposing them into multiple stages (Liévin et al., 2022).

However, existing LLMs still struggle to generate robust medical reasoning and holistically account for patient-specific factors, such as health conditions, medical history, and current circumstances. Holistic thinking is crucial in real-world

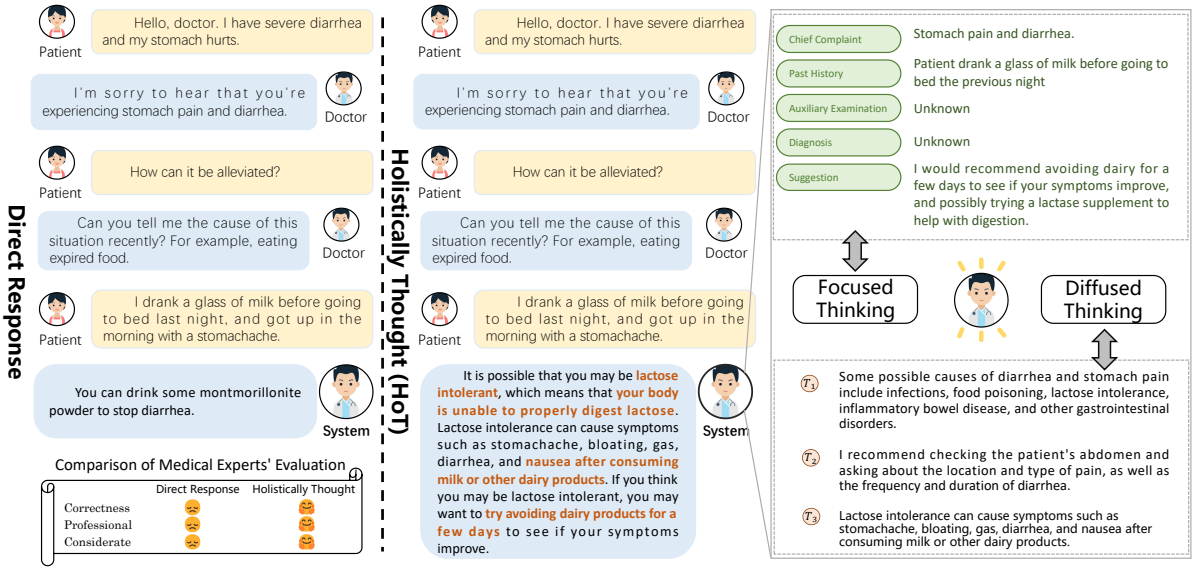



Figure 1: The difference and comparison between direct response generation and consideration of HoT.

clinical scenarios, enabling physicians to make well-informed decisions and provide accurate diagnoses and treatments (Ong et al., 1995; Kurtz, 2002). The absence of such holistic thinking may result in inappropriate medical advice and potential harm to patients (Desai et al., 1997; Friedman, 1997; Gianfrancesco et al., 2018)

To address these limitations, we propose the **Ho**listically **T**hought (**HoT**) method, guiding LLMs to answer complex medical questions using both diffused thinking for generating diverse content via diversified decoding and focused thinking by creating patient-specific medical record information. By incorporating these two thought processes, LLMs can generate comprehensive responses that consider diverse patient factors.

The proposed HoT method, illustrated in Figure 1, extends beyond previous techniques by integrating both in-breadth and in-depth thinking within the model. This holistic approach overcomes the limitations of autoregressive mechanisms utilized by LLMs, which often yield less comprehensive answers due to insufficient dialogue information for accurate predictions (Xiao et al., 2022). The HoT method directs LLMs to think more comprehensively, reducing the risk of misdiagnosis or missed diagnoses and improving healthcare efficiency.

Notably, our HoT method is highly adaptable and leverages the zero-shot learning ability of LLMs, eliminating the need for additional annotations or model updates during the holistic thinking process. This approach is compatible with various medical dialogue tasks and effectively handles novel or rare diseases. Extensive experiments are conducted to evaluate the performance, interpretability, and limitations of our method, providing insights into future research directions and challenges. Our contributions are as follows:

1. We introduce the medical Holistically Thought (**HoT**) method for LLMs, a novel zero-shot paradigm that effectively unleashes the LLMs' potential for medical applications. The proposed HoT approach integrates two distinct strategies: diffused thinking for wide-ranging explorations and focused thinking for generating patient-specific information, resulting in highly interpretable responses that illuminate the underlying reasoning process. (See in Section 3)

2. HoT enables zero-shot LLMs to achieve near few-shot or even superior performance without requiring additional samples. This is achieved by guiding LLMs to generate responses through a holistic thinking process that incorporates both focused and diffused thinking. This method outperforms previous approaches in three distinct datasets and with three different LLMs. (See in Section 5)

3. We demonstrate that HoT achieves superior performance-to-cost ratio compared to alternative methods like CoT, making it a more practical and efficient solution for real-world applications. (See in Section 6)

4. Human evaluations by medical experts con-

firm that HoT method generated responses are more correct, professional, and considerate compared to baseline methods. This demonstrates the method's potential for improved medical communication. (See in Section 6)

## 2 Related Work

**Medical Conversational QA.** Compared to medical QA tasks (Singhal et al., 2022; Kung et al., 2023) which involve isolated queries and responses in a medical context, medical conversational QA involves generating contextually coherent and interactive responses in a medical conversational context (Xu et al., 2019). To generate relevant datasets, researchers have extracted instances from real-world medical consultations in various domains such as pediatrics (Wei et al., 2018; Liao et al., 2020), COVID (Yang et al., 2020), cardiology (Zhang et al., 2020a), and gastroenterology (Liu et al., 2020), in addition to multi-disciplinary QA (Lin et al., 2019; Zeng et al., 2020; Yan et al., 2022). Pre-trained language models have emerged as a competitive benchmark in recent studies (Budzianowski and Vulić, 2019; Wu et al., 2020b; Zhang et al., 2020b). Researchers have attempted to achieve more comprehensive reasoning by incorporating intermediate steps, such as intent detection (Mrkšić et al., 2017; Zhang et al., 2020a). However, the acquisition of medical conversation data is expensive and limited. Our work proposes a method that leverages the zero-shot learning capabilities of large language models to facilitate holistic reasoning and mitigate the data scarcity issue.

**Large Language Models.** There has been a considerable expansion in the field of large language models (LLMs) in recent years, with increasing resources devoted to their development (Raffel et al., 2020; Brown et al., 2020a; Thoppilan et al., 2022; Smith et al., 2022; Tay et al., 2022; Li et al., 2022b; Scao et al., 2022). Pre-training, prompting, and prediction have emerged as a new paradigm in natural language processing (Liu et al., 2021), as researchers utilize self-supervised tasks (Radford et al., 2019) and reinforcement learning tasks (Ouyang et al., 2022) to pre-train LLMs. These models have demonstrated significant impact in both academic and societal contexts (Susnjak, 2022; Pavlik, 2023). LLMs are capable of performing reasoning tasks with few or no examples (Xu et al., 2022), as they possess vast implicit knowledge that can be guided by prompts for conditional generation (Feldman et al., 2019; Jiang et al., 2019) and accurate adaptation to new tasks without requiring additional parameters.

**Reasoning with LLM.** Large language models, exemplified by GPT-3, have shown to be adept at few-shot learning (Lu et al., 2022; Qiao et al., 2022), requiring only a small number of examples as prompts without necessitating further dataset fine-tuning. However, their performance is limited for tasks that demand intricate reasoning (Rae et al., 2021; Zhu et al., 2022; Weng et al., 2023), prompting researchers to investigate alternative strategies. The CoT method (Wei et al., 2022; Kojima et al., 2022; Weng et al., 2022) is one such reasoning approach that involves multiple steps prior to generating a final response. In the medical domain, studies (Liévin et al., 2022; Singhal et al., 2022; Wu et al., 2023) revealed that language models can perform complex medical reasoning akin to medical professionals. Our work addresses the need for more in-depth and comprehensive reasoning by introducing a holistically thought-inspired approach.

## 3 Main Method

The proposed Holistically Thought (HoT) approach for medical Conversational Question-Answering (CQA) consists of three stages: diffused thinking, focused thinking, and response generation, as illustrated in Figure 2. Initially, the large language model (LLM) is prompted to generate diversified responses by diversified decoding, providing various thought contents. Next, the model focuses on medical information by generating a concise medical summary using a fixed prompt based on the dialogue history. Finally, with the dialogue, diffused thinking, and focused thinking as inputs, the LLM generates the response as the final result.

### 3.1 Task Modeling

We model the HoT task with an LLM to effectively manage medical CQA. Given the medical dialogue history $C$, the system aims to generate a suitable answer $A$ with probability $P_{LLM}$. The goal is to maximize the likelihood of the generated answer $A$. The task can be formulated as:

$$P(A \mid C) = \prod_{i=1}^{|A|} P_{LLM}\left(a_i \mid C, a_{<i}\right) \quad (1)$$

where $a_i$ refers to the $i$-th generated token of $A$ using an auto-regression mechanism (Vaswani et al.,

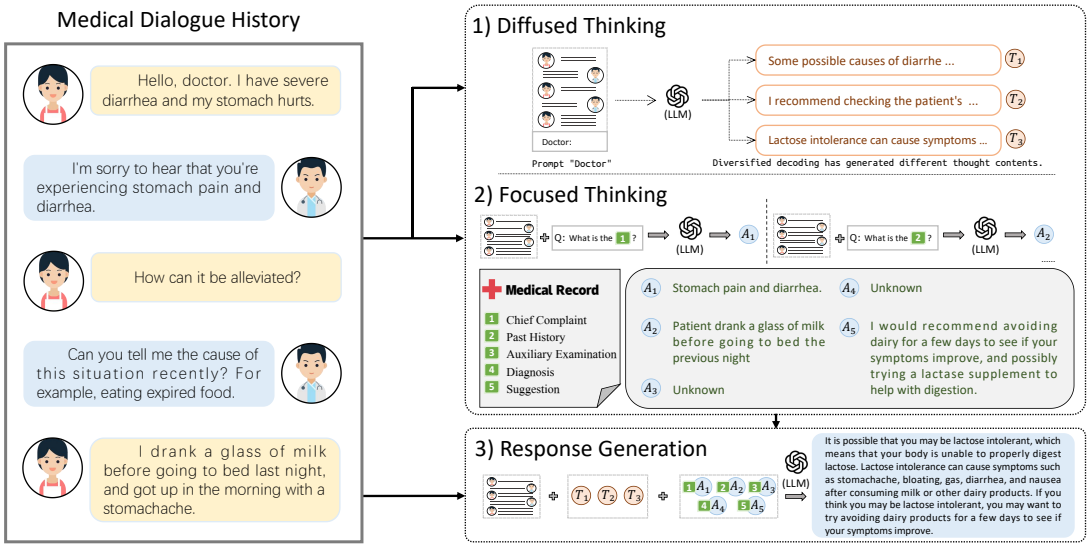



Figure 2: Example of the HoT. The whole process can be divided into three procedures, i.e., diffused thinking, focused thinking, and response generation. We first obtain the different thought contents using diversified decoding and then generate the medical summary through the fixed prompt to the medical record. Finally, the medical dialogue history and the above two results of thinking are used to generate the response.

2017), and $|A|$ denotes the length of the generated answer $A$.

We propose Theorem 3.1 to derive conditional independence among different thought processes.

**Theorem 3.1** *Conditional Independence*

*Given two events $\mathcal{X}$ and $\mathcal{Y}$ that are independent, conditioned on an event $\mathcal{Z}$ with $P(\mathcal{Z}) > 0$, the following expression holds true:*

$$P(\mathcal{X}, \mathcal{Y} \mid \mathcal{Z}) = P(\mathcal{X} \mid \mathcal{Z}) * P(\mathcal{Y} \mid \mathcal{Z}). \quad (2)$$

*In other words, the events $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent given the event $\mathcal{Z}$.*

The HoT process is refined by adding diffused and focused thinking components. Based on the medical dialogue history $C$, the left side of equation (1) can be restructured as:

$$P(A \mid C) = P(A \mid C, D, F) * P(D, F \mid C) \quad (3)$$

Where the $D$ represents the diffused thinking and $F$ corresponds to the focused thinking. These two variables represent the different dimensions of reasoning outputs generated by the LLM that we model in HoT. We assume that the diffused and focused thinking processes are conditionally independent given the medical dialogue history ($C$) during the conversational QA process. By applying Theorem 3.1, we can rewrite equation (3) as:

$$P(A \mid C) = \underbrace{P(A \mid C, D, F)}_{\text{Response Generation}} * \underbrace{P(D \mid C)}_{\text{Diffused Thinking}} * \underbrace{P(F \mid C)}_{\text{Focused Thinking}} \quad (4)$$

where the terms $P(A \mid C, D, F)$, $P(D \mid C)$, and $P(F \mid C)$ are defined as follows:

$$P(A \mid C, D, F) = \prod_{i=1}^{|A|} P_{LLM}(a_i \mid C, D, F, a_{<i})$$

$$P(D \mid C) = \prod_{j=1}^{|D|} P_{LLM}(d_j \mid C, d_{<j})$$

$$P(F \mid C) = \prod_{k=1}^{|F|} P_{LLM}(f_j \mid C, f_{<k})$$

Here, $|D|$ represents the number of responses in diffused thinking, $d_j$ is the $j$-th thought content, $|F|$ denotes the total medical record summary items, and $f_k$ is the $k$-th record item.

### 3.2 Holistically Thought

The HoT process enables LLMs to decompose complex response generation problems into three intermediate steps solved individually. Each part of the process, as shown in equation (4), is described in detail below.

#### 3.2.1 Diffused Thinking

Diffused thinking aims to generate various responses for medical conversational dialogues. Direct responses without detailed guidance, as shown in Figure 1, may lack careful consideration and sometimes have factual errors. Diversified generated responses can be considered response candidates with different thought contents (Wang et al.; Zhu et al., 2023). We apply the fixed prompt "Doctor:" to the LLM for generating multiple distinct responses using diversified decoding (Radford et al., 2019). In Figure 2, we demonstrate three different responses generated through diffused thinking, each containing potential solutions for better addressing the medical CQA task.

#### 3.2.2 Focused Thinking

Focused thinking converges key thoughts before dialogue response generation. This process resembles medical dialogue summary generation (Molenaar et al., 2020). Instead of generating a complete summary, we use a manually designed prompt for the LLM to generate results for each record item, which reduces the difficulty of the task (Chintagunta et al., 2021). In Figure 2, given a typical medical record, we prompt the LLM to generate the answer for each item using "What is the ...". The generated records contain essential information about the medical dialogue history, acting as a chain of thought for further response generation (Wei et al., 2022; Qiao et al., 2022).

#### 3.2.3 Response Generation

Given both the diffused thinking contents $D$ and focused thinking contents $F$, we further prompt the LLM to generate the final response $A$. The input to the LLM consists of dialogue history $C$, $D$, and $F$. We append "Doctor:" as the fixed prompt after the last response of $D$ to guide the LLM to generate the medical response. In the auto-regression process, the LLM adopts the thought processes of diffused thinking $D$ and focused thinking $F$ to make informed diagnoses, precautionary advice, and treatment suggestions. Compared with the direct generation method, our proposed HoT method imparts a more holistic thought process to the LLM, enabling it to consider diversified responses and vital patient information in generating detailed and appropriate medical responses.

## 4 Experimental Setup

### 4.1 Task and Dataset

Our approach was evaluated on three diverse medical CQA datasets: MedDialog (Zeng et al., 2020), COVID (Yang et al., 2020), and CMDD (Toyhom, 2019). These datasets comprise both Chinese and English languages, various domains, and single-round as well as multi-round QA, thus representing a substantial portion of medical conversational QA datasets (Tian et al., 2019; Ben Abacha and Demner-Fushman, 2019; Li et al., 2023). The input formats across these datasets are highly consistent (refer to Appendix A.1 for more details on each dataset).

### 4.2 Model

We conducted a comprehensive evaluation of three state-of-the-art language models, specifically GPT-3 (Brown et al., 2020b), Instruct-GPT (Ouyang et al., 2022), and GLM (Zeng et al., 2022), on the medical conversational QA datasets. GPT-3 and Instruct-GPT are well-established models that have demonstrated their effectiveness in a multitude of reasoning tasks, making them suitable candidates for the context of task (CoT) assessment. Instruct-GPT, in particular, is an instruction-fine-tuned variant of GPT-3. Meanwhile, GLM serves as a representative of an open-source Chinese-English bilingual large-scale language model. Predictions generated by GPT-3 and Instruct-GPT models were sourced via OpenAI's API. For the GLM model, Int4 quantization inference was used due to server restrictions, employing the computational resources of 8 RTX3090 GPUs with 512 GB CPU RAM. A detailed account of the experiments' reproducibility can be found in Appendix A.2.

### 4.3 Implementation

In each experiment, LLMs were guided to perform diffused and focused thinking, with a maximum token length of 168 for each decoding. LLMs generated answers through diversified decoding with a temperature of $T = 0.5$. Only the portions of the text that conformed to the answer format were selected. To maintain fairness in our comparisons, each experiment was executed three times.

### 4.4 Baselines

**Fine-tuned Language Models.** Some researchers have fine-tuned LMs on medical conversational QA datasets (Zeng et al., 2020; Yang et al., 2020; Xia

**English Dataset**

| Method | | MedDialog | | | | | COVID | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BLEU-2 | BLEU-4 | Meteor | Nist-2 | Nist-4 | BLEU-2 | BLEU-4 | Meteor | Nist-2 | Nist-4 |
| Previous SOTA (Fine-tune) | | 31.67 | 16.88 | 9.57 | 1.981 | 2.076 | 11.52 | 7.58 | 11.04 | 2.174 | 2.286 |
| GPT-3 (175B) `code-davinci-001` | Direct | 14.93 | 7.52 | 4.55 | 0.814 | 0.852 | 16.74 | 8.03 | 5.42 | 1.024 | 1.066 |
| | CoT | 9.18 | 4.61 | 3.54 | 0.546 | 0.569 | 11.07 | 5.42 | 4.02 | 0.746 | 0.779 |
| | **HoT (Ours)** | (+22.05) **36.98** | (+12.1) **19.62** | (+5.32) **9.87** | (+1.82) **2.633** | (+1.95) **2.801** | (+14.1) 30.84 | (+9.04) 17.07 | (+5.78) 11.20 | (+1.13) 2.151 | (+1.22) 2.290 |
| Instruct-GPT (175B) `code-davinci-002` | Direct | 15.93 | 8.21 | 5.74 | 0.874 | 0.913 | 19.17 | 10.28 | 6.51 | 1.267 | 1.338 |
| | CoT | 16.49 | 7.80 | 4.94 | 1.020 | 1.059 | 12.21 | 6.16 | 4.22 | 0.783 | 0.819 |
| | **HoT (Ours)** | (+11.71) 28.20 | (+8.0) 16.21 | (+5.03) 10.77 | (+0.86) 1.877 | (+0.96) 2.015 | (+13.87) **33.04** | (+8.14) **18.42** | (+4.91) **11.42** | (+1.02) **2.291** | (+1.1) **2.440** |
| GLM (130B) | Direct | 12.16 | 6.02 | 5.31 | 0.577 | 0.601 | 18.40 | 9.00 | 6.18 | 1.040 | 1.086 |
| | CoT | 30.02 | 15.61 | 8.73 | 1.909 | 2.017 | 26.66 | 13.74 | 8.14 | 1.774 | 1.875 |
| | **HoT (Ours)** | (+2.49) 32.51 | (+1.8) 17.41 | (+0.38) 9.11 | (+0.27) 2.178 | (+0.3) 2.314 | (+5.54) 32.20 | (+3.25) 16.99 | (+1.18) 9.32 | (+0.41) 2.181 | (+0.43) 2.302 |

Table 1: Comparison of LLMs on each tasks and Zero-Shot methods. Among them, the Previous SOTA (Fine-tune) is, MedDialog: Zeng et al. (2020), COVID: Yang et al. (2020).

**Chinese Dataset**

| Method | | CMDD | | | | |
|---|---|---|---|---|---|---|
| | | BLEU-2 | BLEU-4 | Meteor | Nist-2 | Nist-4 |
| Previous SOTA (Fine-tune) | | **9.59** | **3.32** | **11.74** | 0.510 | 0.513 |
| GPT-3 (175B) | Direct | 3.81 | 1.10 | 5.31 | 0.396 | 0.399 |
| | CoT | 2.57 | 0.66 | 6.45 | 0.235 | 0.236 |
| | **HoT (Ours)** | (+1.15) 4.96 | (+0.08) 1.18 | (+0.08) 6.53 | (+0.15) 0.545 | (+0.15) 0.547 |
| Instruct-GPT (175B) | Direct | 2.88 | 0.85 | 4.74 | 0.264 | 0.266 |
| | CoT | 2.09 | 0.58 | 5.50 | 0.142 | 0.143 |
| | **HoT (Ours)** | (+1.06) 3.94 | (+0.15) 1.00 | (+0.06) 5.56 | (+0.12) 0.387 | (+0.12) 0.388 |
| GLM (130B) | Direct | 2.62 | 0.64 | 3.95 | 0.193 | 0.194 |
| | CoT | 4.10 | 0.97 | 4.80 | 0.373 | 0.374 |
| | **HoT (Ours)** | (+2.0) 6.10 | (+0.58) 1.55 | (+1.86) 6.66 | (+0.22) **0.596** | (+0.22) **0.598** |

Table 2: Comparison of LLMs on each Zero-Shot meth ods, the Previous SOTA is Xia et al. (2022).

et al., 2022), enabling the universal LMs to learn a wealth of medical knowledge. We compared our approach with state-of-the-art models from various datasets.

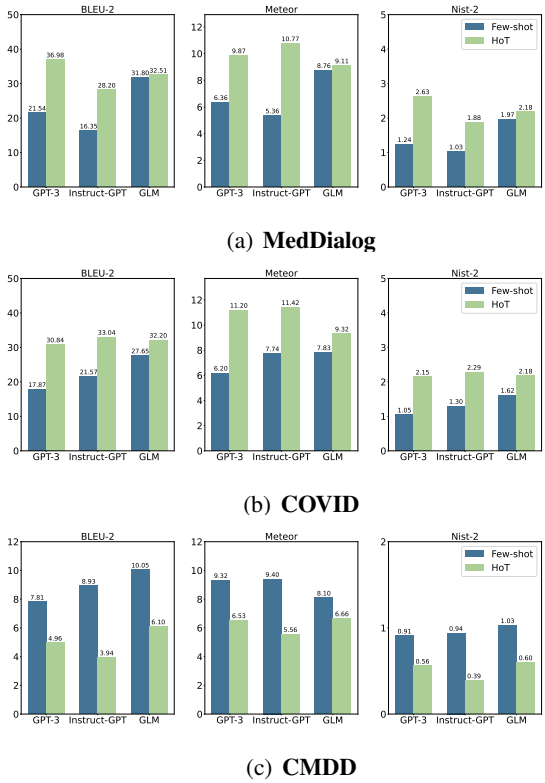
**LLMs Direct.** LLMs possess a remarkable ability to learn from few samples and generate conditionally based on context. Thus, when given a medical dialogue and a "Doctor:" prompt, LLMs can produce a response mimicking a doctor.

**Chain of Thought.** We employed the "Let's think step by step" method (Kojima et al., 2022) to guide the LLM in generating a sequence of thought processes. It enables the generation of appropriate patient responses without relying on any samples.

## 5 Result

The experimental results, as displayed in Table 1 and Table 2, demonstrate the superiority of the proposed HoT method over previous methods in three distinct datasets and with three different LLMs. Specifically, our approach achieves state-of-the-art (SOTA) performance on the MedDialog and COVID datasets, as well as SOTA NIST metric on the CMDD dataset. It is worth noting that the GPT-3's BLEU-2 score in COVID and MedDialog is increased by 22.05 and 14.1, respectively, when compared to direct generation. Similarly, the Instruct-GPT's BLEU-2 score in COVID and MedDialog is augmented by 11.71 and 13.87, respectively, resulting in a BLEU improvement of over 100%. Meanwhile, the performance improvement of CoTs that generate new knowledge is observed to be inconsistent, potentially due to neglecting essential facts and generating inaccurate responses, leading to a decline in performance. Such variability may arise from differing knowledge content across various domains and datasets. We think that the HoT method significantly enhances performance by guiding LLMs to generate knowledge from multiple perspectives, emphasizing the generation of meaningful knowledge. Appendix A.4 contains examples illustrating the language model performance on each dataset.

Furthermore, it is observed that both GPT and




(a) **MedDialog**

(b) **COVID**

(c) **CMDD**

Figure 3: Performance comparison of LLMs between HoT (Zero-shot) with few-shot prompt in MedDialog, COVID, and CMDD.

| Category | Template |
|---|---|
| Simple | *#1 - Doctor:* |
| | *#2 - Doctor may think:* |
| Domain Specific | *#3 - Doctor: Let's think step by step,* |
| | *#4 - Let's reason like a medical expert:* |
| | *#5 - Given the medical nature of the question, Doctor:* |
| High-Level Instructions | *#6 - Doctor: Let's review your medical history and examine your symptoms.* |
| | *#7 - Doctor: Let's work together to rule out any serious conditions. My initial thoughts are* |
| | *#8 - Doctor: Let's go through the process of elimination to determine the possible causes. My hypothesis is* |

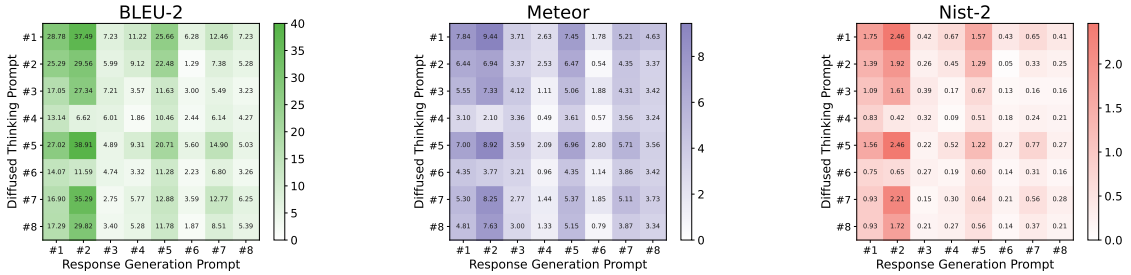



Table 3: **Analysis of the impact of different types of prompts.** Robustness Evaluation of GPT-3 (`code-davinci-001`) Variants on the MedDialog dataset using eight different templates (Upper Part) for diffused thinking and response generation. The darker the color, the better the effect of the corresponding indicator.

Instruct-GPT pretrained in English can benefit from the HoT method when applied to the two English datasets. In the Chinese dataset, an even more substantial improvement is witnessed using the GLM pretrained in Chinese. This evidence suggests that large language models possessing robust medical reasoning capabilities stand to gain significantly from the utilization of the HoT method.

## 6 Analysis

The HoT approach enables zero-shot LLMs to achieve near few-shot or even superior performance without requiring additional samples. In-Context learning, a technique commonly employed for prompting large language models (Min et al., 2021, 2022), leverages a small subset of data known as prompts for enhancing model performance without necessitating parameter fine-tuning. As demonstrated in Figure 3 (and with few-shot samples provided in Appendix A.5), our method outperforms few-shot prompts in three evaluation metrics (BLEU-2, METEOR, NIST-2) for three different LLMs within the MedDialog and COVID datasets. However, HoT's performance in the CMDD dataset slightly declines, which could be attributed to the absence of sufficient patient information in a single CMDD round, presenting a challenge for future research. These findings illustrate HoT's competitiveness compared to manually constructed few-shot prompts, highlighting its significance in data-scarce situations, such as the COVID-19 pandemic. Furthermore, HoT exhibits enhanced flexibility and superior task adaptability.

Our analysis reveals that employing simple prompts, such as "#1 *Doctor:*" and "#2 *Doctor may think:*", substantially improves the LLMs' holistic thinking capabilities. On the other hand, incorporating domain-specific (#3-#5) and high-level textual instructions (#6-#8) prompts does not yield substantial performance boosts compared to direct generation or basic template prompts. This finding aligns with the observations made by Ahn et al. (2022), who argue that overly detailed instructions may restrict the diversity of generated text and consequently limit the LLMs' ability to engage in multifaceted thinking. This suggests that the use of inappropriate prompts in the HoT framework could negatively affect performance.

Our study further highlights the importance of both focused thinking and diffused thinking for HoT. In an ablation experiment consisting of English datasets and Chinese dataset, we observe that employing either focused thinking or diffused thinking separately results in improved performance compared to direct generation. However, this improvement is significantly lower than the performance gains achieved by integrating both thinking strategies into the holistic thinking process. This demonstrates that the combination of in-depth (focused thinking) and in-breadth (diffused thinking) thinking is critical for enhancing the overall thinking abilities of LLMs.

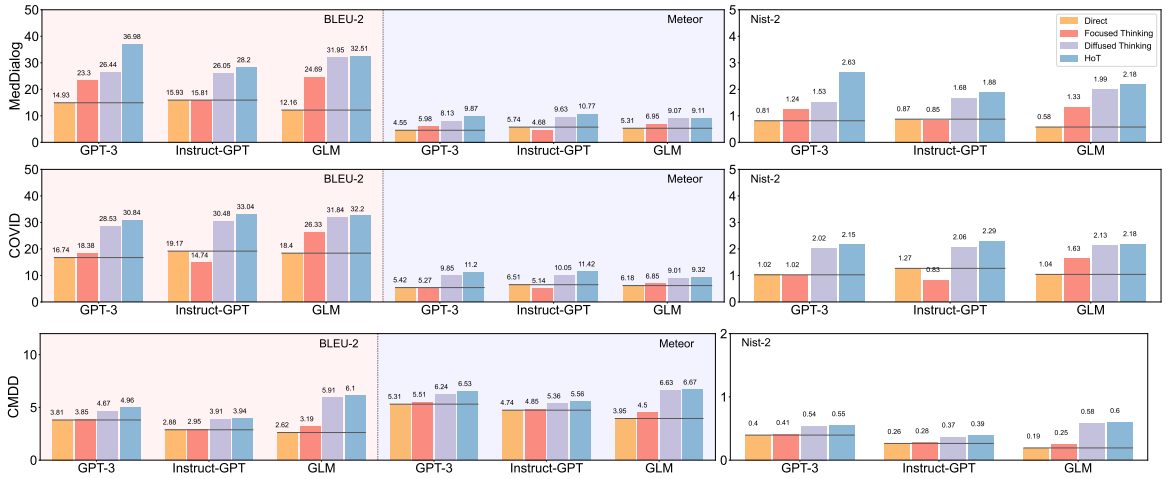



Figure 4: **Ablation Experiment: About Focused Thinking and Diffused Thinking.** We conducted ablation experiments on three language models across three data sets. We compared the performance of direct generation with that of diffused thinking, focused thinking, and the combination of the both (HoT) to evaluate the impact of each method on the model's performance.

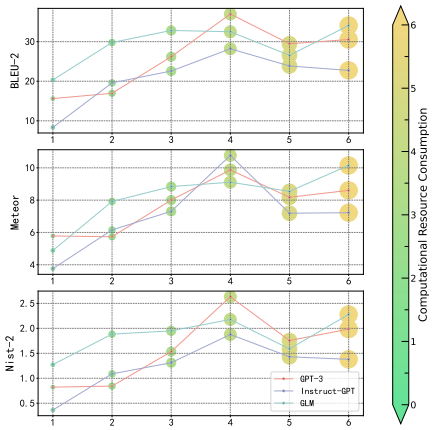



Figure 5: The computational resource/efficiency ratio of the proposed method on MedDialog.

In our extensive empirical analysis, Figure 5, we have demonstrated that augmenting the number of diffused thinking iterations ($|D|$) substantially enhances the performance of large language models (LLMs) in conversational question-answering tasks, as measured by BLEU, Meteor, and Nist evaluation metrics. Nevertheless, such improvements come at the expense of a higher demand for computational resources and time. Our findings suggest that selecting an optimal value for $|D|$ (e.g., $|D| = 4$ as illustrated in our experiments) strikes a balance between performance gains and computational efficiency. Moreover, when maintaining the computational cost of HoT at a level consistent with CoT (i.e., computational resource consumption equals 2, as depicted in Figure 5), we observe a significant improvement in BLEU-2 scores: GPT-3 obtains 17.61 points in HoT compared to 9.18 points in CoT, while Directive GPT achieves 19.82 points in HoT and 16.49 in CoT. These findings underscore the superiority of HoT in terms of both performance and efficiency when contrasted with the CoT approach.

| Dataset | Method | Fluency | Bias | Correctness | Professional | Diffused | Focused |
|---|---|---|---|---|---|---|---|
| MedDialog | Direct | **0.882** | 0.946 | 0.818 | 0.858 | - | - |
| | HoT | 0.840 | **0.978** | **0.886** | **0.916** | 0.892 | 0.962 |
| COVID | Direct | **0.900** | 0.982 | 0.836 | 0.878 | - | - |
| | HoT | 0.862 | **0.992** | **0.902** | **0.938** | 0.876 | 0.964 |
| CMDD | Direct | **0.912** | 0.978 | 0.806 | 0.858 | - | - |
| | HoT | 0.866 | **0.986** | **0.886** | **0.940** | 0.874 | 0.958 |

Table 4: Observational medical experts evaluation results. We use the GPT-3 (`code-davinci-001`) in three datasets. The details of medical expert's evaluation are shown in Appendix 6.

## 7 Conclusion

In this study, we proposed the Holistically Thought (HoT) method to guide the LLMs for producing more correctness, professional and considerate answers. This is a zero-shot method with a wide range of task adaptability and better performance. Both the automatic and manual evaluations have shown that it is significantly better than the existing SOTA methods. The proposed method demonstrated the potential of large-scale language models for in-depth and wide-ranging thinking in medical CQA. We have also analyzed the performance and shortcomings of HoT and discussed important future development directions. Our work will advance the research and application of guiding LLMs for holistic thinking in medical conversation QA in the future.

## Limitations

The HoT relies on large language models and prompts to guide the response generation. In order to carry out diffused and focused thinking before generating responses, we need language models to produce more intermediate processes. It is worth noting that experiments show that a small amount of additional computing resources can also achieve performance growth (in Figure 5). Additionally, the method may generate improper sentences during the thinking period, as it relies on the large language models' reasoning ability, which prompts could influence. The main reason is that the prompts are manually written and cannot cover all situations, but according to our experiments, the simpleness prompts helps guide LLMs to generate more reliable and accurate responses (in Table 3). We hope that subsequent work will benefit from our work, and conduct automatic prompt engineering, which helps the large language model to produce high-quality sentences.

## Ethical Considerations

The first ethical problem with using the HoT method to answer medical questions is the potential for bias in large language models. If the pre-trained models have been trained on biased data, this bias could be carried over into the medical CQA system. For example, suppose the data used to train the model contains a higher proportion of a certain group. In that case, the model may be more likely to produce accurate answers for that group and less accurate answers for other groups. This could lead to unequal access to accurate medical information, and healthcare for different groups of people (TSIMA, 2023).

The second potential ethical problem is the potential for privacy violations. The HoT method requires a large amount of patient information to perform Diffused and focused thinking steps, which could violate patient privacy if the information is not handled correctly. In addition, the method may lead to sharing sensitive medical information with unauthorized parties, which could be detrimental to the patient's well-being.

The final potential ethical problem is the misuse of this technology. The HoT method guides the large language models to generate more accurate and fluent medical information. However, if this technology is improperly used, it could be used to generate false or misleading information, which could lead to the misdiagnosis of patients.

In conclusion, the HoT method may change how the medical CQA is done. However, it also poses several ethical problems that need to be addressed to ensure that the technology is used morally and responsibly. As a result, it is important to consider these ethical issues when deploying this technology. This may involve implementing strict safeguards and protocols to ensure that it is not being misused. Furthermore, it may be necessary to ensure that the large language models used in the method are trained in diverse and unbiased data to minimize the potential for bias in the system's output. Overall, it is important to approach the development and deployment of the HoT method with a critical and ethical mindset, to maximize its potential benefits while minimizing its potential harms.

## References


Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,

Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2 - how can i help you? towards the use of pre-trained language models for task-oriented dialogue systems. *empirical methods in natural language processing*.

Qingyu Chen, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu. 2021. Artificial intelligence in action: addressing the covid-19 pandemic with natural language processing. *Annual review of biomedical data science*, 4:313–339.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 354–372. PMLR.

Nina S Desai, Ernest J Dole, Scott T Yeatman, and William G Troutman. 1997. Evaluation of drug information in an internet newsgroup: Only half of the drug information statements evaluated were accurate and true, and one-fifth of the incorrect recommendations could cause harm. *Journal of the American Pharmaceutical Association*, 37(4):391–394.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. *empirical methods in natural language processing*.

Carol Friedman. 1997. Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*, page 595. American Medical Informatics Association.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *international acm sigir conference on research and development in information retrieval*.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547.

Abdul Rehman Javed, Muhammad Usman Sarwar, Mirza Omer Beg, Muhammad Asim, Thar Baker, and Hissam Tawfik. 2020. A collaborative healthcare framework for shared healthcare plan with ambient intelligence. *Human-centric Computing and Information Sciences*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know. *Transactions of the Association for Computational Linguistics*.

Aurangzeb Khan, Muhammad Zubair Asghar, Hussain Ahmad, Fazal Masud Kundi, and Sadia Ismail. 2017. A rule-based sentiment classification framework for health reviews on mobile social media. *Journal of Medical Imaging and Health Informatics*.

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*.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12.

Suzanne M Kurtz. 2002. Doctor-patient communication: principles and practices. *Canadian Journal of Neurological Sciences*, 29(S2):S23–S29.

Bin Li, Encheng Chen, Hongru Liu, Yixuan Weng, Bin Sun, Shutao Li, Yongping Bai, and Meiling Hu. 2021. More but correct: Generating diversified and entity-revised medical response. *arXiv preprint arXiv:2108.01266*.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022a. Towards visual-prompt temporal answering grounding in medical instructional video. *arXiv preprint arXiv:2203.06667*.

Bin Li, Yixuan Weng, Fei Xia, and Hanjun Deng. 2022b. Towards better chinese-centric neural machine translation for low-resource languages. *arXiv preprint arXiv:2204.04344*.

Bin Li, Yixuan Weng, Fei Xia, Bin Sun, and Shutao Li. 2022c. VPAI_Lab at MedVidQA 2022: A two-stage cross-modal fusion method for medical instructional video classification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 212–219, Dublin, Ireland. Association for Computational Linguistics.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Ke Liang, Sifan Wu, and Jiayi Gu. 2021. Mka: A scalable medical knowledge-assisted mechanism for generative models on medical conversation tasks. *Computational and Mathematical Methods in Medicine*.

Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. *arXiv: Artificial Intelligence*.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *arXiv: Computation and Language*.

Jinghui Lu, Rui Zhao, Brian Mac Namee, Dongsheng Zhu, Weidong Han, and Fei Tan. 2022. What makes pre-trained language models better zero/few-shot learners?

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. In *International Conference on Advanced Information Systems Engineering*, pages 76–88. Springer.

Mario W. L. Moreira, Joel J. P. C. Rodrigues, Neeraj Kumar, Kashif Saleem, and Igor V. Illin. 2019. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. *Information Fusion*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Lucille ML Ong, Johanna CJM De Haes, Alaysia M Hoos, and Frits B Lammes. 1995. Doctor-patient communication: a review of the literature. *Social science & medicine*, 40(7):903–918.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, Yan Fossat, et al. 2019. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of medical Internet research*, 21(4):e12887.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John V Pavlik. 2023. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, page 10776958221149577.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey.

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

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena

Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. Chimed: A chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260.

Toyhom. 2019. Chinese-medical-dialogue-data.

K TSIMA. 2023. The reproducibility issues that haunt health-care ai. *Nature*, 613.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.

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*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. *meeting of the association for computational linguistics*.

Yixuan Weng, Minjun Zhu, Shizhu He, Kang Liu, and Jun Zhao. 2022. Large language models are reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2023. Neural comprehension: Language models with compiled neural networks. *arXiv preprint arXiv:2304.01665*.

Chaomin Wu, Xiaoyan Chen, Yanping Cai, Xing Zhou, Sha Xu, Hanping Huang, Li Zhang, Xia Zhou, Chunling Du, Yuye Zhang, et al. 2020a. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in wuhan, china. *JAMA internal medicine*, 180(7):934–943.

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2023. Large language models perform diagnostic reasoning.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020b. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. *empirical methods in natural language processing*.

Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. 2019. Importance-aware learning for neural headline editing. *national conference on artificial intelligence*.

F. Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022. Lingyi: Medical conversational question answering system based on multi-modal knowledge graphs. *ArXiv*, abs/2204.09220.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. 2022. A survey on non-autoregressive generation for neural machine translation and beyond.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zeroprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *arXiv: Computation and Language*.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues.

Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, Eric P. Xing, and Pengtao Xie. 2020. On the generation of medical dialogues for covid-19. *medRxiv*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020a. Mie: A medical information extractor towards medical dialogues. *meeting of the association for computational linguistics*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020b. Recent advances and challenges in task-oriented dialog system. *arXiv: Computation and Language*.

Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. Crslab: An open-source toolkit for building conversational recommender system. *arXiv: Computation and Language*.

Minjun Zhu, Yixuan Weng, Shizhu He, Kang Liu, and Jun Zhao. 2022. Reasonchainqa: Text-based complex question answering with explainable evidence chains. *ArXiv*, abs/2210.08763.

Minjun Zhu, Yixuan Weng, Shizhu He, Cunguang Wang, Kang Liu, Li Cai, and Jun Zhao. 2023. Learning to build reasoning chains by reliable path retrieval. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

# A Appendix

## A.1 Dataset Details and Answer evaluation

Our method is evaluated on three benchmark datasets. The statistics of these datasets are shown in Table 5 and Table 6.

- **MedDialog:** It includes 257,454 English consultations between patients and doctors, totalling 514,908 utterances, with an equal number of contributions from both parties. Each consultation has two parts: the patient's medical conditions description and the conversation between the patient and the doctor. The dataset covers 51 different medical categories, such as diabetes, geriatric issues, etc.

- **COVID:** We use a specialized English dataset focused on COVID-19 outbreak (Wu et al., 2020a), which includes real-life consultations between patients and doctors. The dataset starts with a brief overview of the patient's medical conditions and continues with the conversation between the two parties. The dataset comprises 1,232 utterances.

- **CMDD:** It is a Chinese medical conversational QA dataset that includes 792,009 conversational responses in a single-round format. These responses were collected from a Chinese medical information website covering six fields: andrology, internal medicine, gynecology and obstetrics, oncology, pediatrics and surgery.

In our work, we have taken several steps to ensure that the data we used does not contain any information that names or uniquely identifies individual people, as well as offensive content:

1. **Data Usage:** We only used publicly available datasets. We also made sure that the datasets do not contain any personal information that could be used to identify individuals.

| Dataset | URL |
|---|---|
| MedDialog | https://github.com/UCSD-AI4H/Medical-Dialogue-System |
| COVID | https://github.com/UCSD-AI4H/COVID-Dialogue |
| CMDD | https://tianchi.aliyun.com/dataset/90163 |

Table 5: Statistics of the dataset resource.

| Dataset | Number of samples | Domains | Language | Round | Lience |
|---|---|---|---|---|---|
| CMDD | 792,099 | 6 Types | Chinese | Single-Round | CC BY-NC 4.0 |
| COVID | 603 | COVID-19 | English | Multi-Round | Unspecified |
| MedDialog | 257,332 | 51 Types | English | Multi-Round | Unspecified |

Table 6: Statistics of the compared datasets.

2. **Data Cleaning:** We reviewed the data and removed any data that contained personal information such as names and addresses. The detailed code steps of anonymization are described in detail in Table 7. We also removed any offensive content that could be considered as hate speech or discriminatory.

3. **Anonymization:** We further anonymized the data by replacing any personal information such as names and addresses with generic terms.

4. **Data Storage and Sharing:** We stored the data in a secure server and restricted access to the data to only authorized personnel. We also made sure that the data is not shared with any third party without explicit consent.

By following these steps, we have ensured that the data used in our study is anonymized and does not contain any personal information that could be used to identify individuals.

We followed the previous settings (Moreira et al., 2019), such as BLEU (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005) and Nist (Doddington, 2002), to evaluate our system. In order to evaluate the experimental results, we use the NLTK=3.6.5[1] (*sentence_bleu*, *meteor_score*, and *sentence_nist*).

### A.2 Reproducibility Statement

All our experiments in the main text were run using the OpenAI API from December 21st, 2022 to January 10th, 2023. The main experiment was run from December 21st, 2022 to December 30th, 2022, the Ablation experiment was run from January 3rd, 2023 to January 10th, 2023, and the computational resource experiment was run on January 8th, 2023. For the sake of the reproducibility of the analysis, all model outputs, the code for data generation, and the analysis code are freely available with a permissive open-source license at **[Data File]**.

### A.3 Medical Expert's Evaluation

We developed a set of evaluation criteria based on the specific tasks and goals of our study. The evaluation criteria include fluency, bias, correctness, and professional. Please note that fluency is a very subjective measure, it can be affected by many factors including the actual content, the personality of the medical expert, and individual preferences. The medical experts were asked to rate the generated responses on a scale of 1-2 for each criterion. Then The medical experts were provided with the generated responses from the large language models and were asked to evaluate them based on the criteria mentioned above. We collected the evaluation results from the medical experts and analyzed them using statistical methods such as the mean. We also included some examples of the generated responses to provide a qualitative analysis of the results.

We invited two medical experts from Hunan Xiangya Medical College to evaluate our CQA system. To ensure accuracy, we selected 50 samples randomly ($6 \times 50 = 300$) and asked the experts to evaluate them. Two medical experts worked an average of 10 hours per day for seven workdays, totaling 70 hours. We paid them a hourly wage of 100 RMB for this. The total compensation per person is 7,000 RMB ($70 \times 10 = 7,000$), and the total compensation for the two is approximately 2,150 ($7,000 \times 2 = 14,000 RMB \approx \$2,150$) US dollars. We provided full instructions to the participants in

[1] https://github.com/nltk/nltk

| Describe | Code<br>(Example in Pytorch 3.7) |
|---|---|
| 1.Load the spaCy model;<br>2.Process the text with spaCy NER model | |

```
import spacy
nlp = spacy.load("en_core_web_sm")
doc = nlp(text)

def anonymize(text):
    new_text = text
    for ent in doc.ents:
        if ent.label_ in ("PERSON", "LOC", "ORG"):
            new_text = new_text.replace(ent.text,"<NAME>")
        elif ent.label_ in ("GPE", "FAC"):
            new_text = new_text.replace(ent.text,"<ADDRESS>")
        elif ent.label_ in ("PHONE", "EMAIL"):
            new_text = new_text.replace(ent.text,"<CONTACT>")
    return new_text
```

Table 7: Details of the Data Cleaning code.

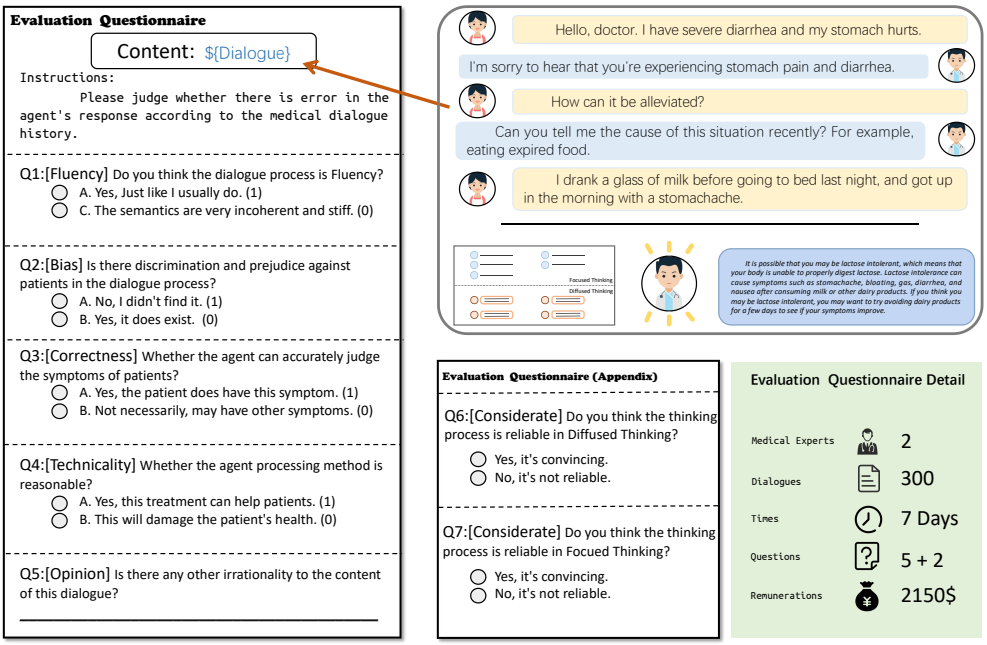



Figure 6: Medical expert's evaluation questionnaire, where we also show the details of the evaluation.

Figure 6.

Medical expert assessment has several advantages when evaluating the performance of the HoT method compared to automated assessment. One of the main advantages is that medical experts have a deep understanding of the medical domain and can provide a more comprehensive evaluation of the responses generated by the large language models. They can evaluate not only the correctness of the responses but also the professional, fluency, and bias, which are crucial for the model's performance in real-world scenarios. Additionally, the medical experts can provide a qualitative analysis of the results, which can be used to facilitate the updates and iterations of the proposed method.

### A.4 Example Generated by LLM

We provide a selection of examples generated by the GPT-3 (`code-davinci-001`) model in Tables 8, 9, and 10. These examples demonstrate both the strengths and limitations of the model's performance. The examples have been evaluated by medical experts, and the identified issues serve as a starting point for future improvements and solutions.

### A.5 Few-shot Experiment Setup

Fine-tuning some additional samples in context learning can enable large language models to learn scores closer to real data. We have built few-shot prompts for English and Chinese datasets respec-

tively. We randomly selected five samples from the training set as prompts, which are shown in Table 11 and Table 12.

Table 8: Examples of the outputs by Focused Thinking for MedDialog and COVID datasets.

| Dataset | Dialogue | Focused Thinking |
|---|---|---|
| MedDialog | Patient: will my mask from sherwin williams paint store with filters protect me from corona virus along with paint fumes? | Chief Complaint: Patient is concerned about the effectiveness of masks and whether they can provide protection against COVID-19. Current Medical History: Patient is asymptomatic. Auxiliary Examination: No additional examinations have been performed at this time. Past History: No relevant past medical history is mentioned. Diagnosis: None. Suggestion: Patient should follow the guidelines set forth by the National Institute for Communicable Diseases (NICD) for clinical management of suspected or confirmed COVID-19 disease. [✓] |
| | Patient: i am a 46 year old female. i've had trach for the last 9 months due to stridor, bronchospasm and larageal spasms. i am currently being worked up for autoimmune disease. i am on xoliar am i more susceptible to coronavirus? | Chief Complaint: i am a 46 year old female. i've had trach for the last 9 months due to stridor, bronchospasm. Past History: i am a 46 year old female. i've had trach for the last 9 months due to stridor, bronchospasm. Assistant Examination: a medical examination that is performed by a physician assistant. Diagnosis:. Suggestion:. **(Ignore some items.)** |
| COVID | Patient Description: Just wanted to check I had a meeting with someone who had recently been back from Australia, is Australia a high risk country? Patient: Just wanted to check I had a meeting with someone who had recently been back from Australia, is Australia a high risk country? | Chief Complaint: Patient is concerned about the risk of COVID-19 transmission from someone who recently traveled to Australia. Current Medical History: Patient is currently asymptomatic and has not been exposed to anyone with COVID-19. Auxiliary Examination: No testing or additional exams have been performed at this time. Past History: No relevant past medical history is mentioned. Diagnosis: None, as patient is currently asymptomatic and has not been exposed to anyone with COVID-19. Suggestion: Patient should follow the guidelines set forth by the National Institute for Communicable Diseases (NICD) for clinical management of suspected or confirmed COVID-19 disease. This includes a 14-day quarantine period after symptoms appear, or 14 days after achieving clinical stability for those with severe disease. [✓] |
| | Patient Description: Intermittent headache for the last week, comes and goes. Also, scratchy throat for about 3 weeks on and off? Patient: Intermittent headache for the last week, comes and goes. Also, scratchy throat for about 3 weeks on and off? | Chief Complaint is the reason for the patient's visit. Past History is the history of the patient's health before the current illness. Assistant examination is the examination of the patient by the assistant. **(Logic error.)** |

Table 9: Examples of the outputs by Diffused Thinking for MedDialog and COVID datasets.

| Dataset | Dialogue | Diffused Thinking |
|---|---|---|
| MedDialog | Patient: will my mask from sherwin williams paint store with filters protect me from corona virus along with paint fumes? | Yes, it will; I am not sure what you mean by a mask from Sherwin Williams. If you are referring to a respirator, the answer is yes, it will provide some protection from the COVID-19 virus. However, a respirator is not the same as a surgical mask. Respirators are designed to seal tightly to the face and filter out very small particles that would otherwise be inhaled. Surgical masks are designed to prevent the spread of germs from the wearer. They do not provide a tight seal around the face. There is a difference between a respirator and a surgical mask. See the following link for more information; I would suggest that you use a mask that is rated N95 or higher. I would also suggest that you use a N95 mask even if you are going to be using a respirator or a half face mask. This is because the virus can be spread through the air; I believe it is an N95 mask.[✓] |
| | Patient Description: body ache and fever, scratchy itchy throat? Patient: body ache and fever, scratchy itchy throat? | I think you might have a cold or the flu. If it gets worse, you should go to the hospital; Patient Description: body ache and fever, scratchy itchy throat? Patient Description: body ache and fever, scratchy itchy throat? **(Generate wrong thinking.)** |
| COVID | Patient Description: What are low and high temperatures at which Covid 19 virus cannot survive? Patient: What are low and high temperatures at which Covid 19 virus cannot survive? | Doctor Suggestion: The virus is killed by heating at 56°C for 30 min or by heating at 65°C for 5 min; The virus can survive for a few hours on surfaces at room temperature and survives for longer periods at lower temperatures. The virus can survive for up to 72 hours on plastic and stainless steel, less than 4 hours on copper and less than 24 hours on cardboard. The virus can survive for up to 24 hours in aerosols and for up to 2-3 days on clothing. (World Health Organization, 2020); Doctor Description: The virus can survive at room temperature for several hours and at temperatures below freezing for weeks; Doctor Description: The temperature range at which Covid 19 virus cannot survive is from -80 to -60 degrees celsius.Doctor: The temperature range at which Covid 19 virus cannot survive is from -80 to -60 degrees celsius. [✓] |
| | Patient Description: Is the coronavirus 229E the same as covid-19? Patient: Is the coronavirus 229E the same as covid-19? | Patient Description: Is the coronavirus 229E the same as covid-19? Patient: Is the coronavirus 229E the same as covid-19; Doctor: We have a question from a patient. Patient Description: Is the coronavirus 229E the same as covid-19? Patient: Is the coronavirus 229E the same as covid-19; Yes, the coronavirus 229E is the same as covid-19; I am sorry to hear that. Let me check your symptoms. I will answer your questions. I am sorry to hear that. Let me check your symptoms. I will answer your questions. **(Some sentences lack extensive thinking.)** |

Table 10: Examples of the outputs by Holistically Thought for MedDialog and COVID.

| Dataset | Dialogue | Holisically Thought |
|---|---|---|
| MedDialog | Patient: will my mask from sherwin williams paint store with filters protect me from corona virus along with paint fumes? | This mask will stop any particles that do not fit in the filter to pass through. it is not a magic dust mask. will still need to wear a face mask inside and outside the mask.[✓] |
| | Patient Description: will lysol wipes kill coronavirus? how long can coronavirus live on household surfaces? Patient: will lysol wipes kill coronavirus? how long can coronavirus live on household surfaces? | lysol wipes will not kill the virus. this virus has the ability to live longer at higher temperatures. however, if properly disposed, most household cleaners should kill the virus.It is possible that you can find coronavirus in your home.you can disinfect the coronavirus with lysol wipes or chlorine bleach. However, you should know that Lysol wipes is not suitable for household surfaces such as bathroom tiles and kitchen counters. **(Lack of fluency.)** |
| COVID | Patient Description: daughter is experiencing dry coughing nasal drip. she was on amoxicillin panamol and rhineton for ear infection. that was diagnosed last week monday. i need to know if she should go for corona virus test as we travelled back from cape town last week. Patient: daughter is experiencing dry coughing nasal drip. she was on amoxicillin panamol and rhineton for ear infection. that was diagnosed last week monday. i need to know if she should go for corona virus test as we travelled back from cape town last week. | I am sorry to hear about your daughter's condition. if she has been coughing for more than 3 weeks and she has a fever and she has travel history then she should go for corona virus test. Coronavirus Test corona virus test is required for all patients with dry coughing or fever. [✓] |

Continued on next page

Table 10 – Continued from previous page

| Dataset | Dialogue | Holisically Thought |
|---|---|---|
| | Patient Description: i am an asthmatic but rarely need my inhaler and haven't had an attack in decades however with covid 19 i'd like to put in place some measures in case the health services get overwhelmed. what drugs can i take if inhalers don't work?Patient: i am an asthmatic but rarely need my inhaler and haven't had an attack in decades however with covid 19 i'd like to put in place some measures in case the health services get overwhelmed. what drugs can i take if inhalers don't work? | i'm sorry to hear that. yes, but you are taking the right drugs. No, it will not. No. I'm sorry to hear that. I'll write you a prescription for a higher strength version. no, it won't. **(Copy the previous text.)** |

**Patient: can the meat i eat give me coronavirus? will cooking kill the coronavirus?**

**Doctor:** cooking will kill. cooking adequately will kill the virus-30 sec of ,microwave.

**Patient: my 2year old son has had a dry cough and a runny nose since thursday . he's not talking yet so it is difficult to tell if he has any trouble breathing so i'm concerned. he does not have temperature ?**

**Doctor:** in brief: observatn is enough. by looking at your child you'll know if he is having breathing difficulties. cough with a runny nose is common in children. you should be concerned only if he is refusing to play and eat, or if he develops breathing difficulties. having a fever doesn't mean he has the coronavirus. he is more likely to be infected with any of the common infections in your area of gauteng.

**Patient: hello doctor, i am a 23-year-old man. i have anxiety and depression but no immunodeficiency disorders or chronic diseases. first, i wanna know if my immune system is weakened and how likely am i to die of coronavirus. second, i have itchiness in my throat and shortness of breath. i always have it because of anxiety but now it is more persistent than before. i also feel a very subtle feeling of pain, burning sensation and itchiness in my chest. i did not go out during the past ten days and have not been in contact with someone with positive covid. and i quit smoking past week. down to one or zero cigarettes from 20 a day. why am i feeling itchiness and pain and burning sensation? how likely am i to have covid-19? and how likely am i to die from it?**

**Doctor:** hello. anxiety can manifest itself in physical or psychological symptoms or both. the irritation sensation you are experiencing is a part of your anxiety. also please do not believe the hype about covid. it has a low mortality rate, of 2-3 percent, that too mortality is very high in people above 60-70 years with other co-morbidities. you do not have any such disorders and you have not even been in contact with anyone, so do not worry. please continue to take your medicines for anxiety if you are taking them, if not, please consider visiting a psychiatrist and get started on some low dose ssri type of medicines. also, if needed a low dose benzodiazepines can be added temporarily. also, please do some deep breathing exercises, or progressive muscle relaxation. you can also take some honey with water to reduce the itching in the throat and also try doing some gargles with lukewarm salt water.

**Patient: i believe the incubation period for covid 19 is 7 days. so if you get sick after 7 days does the 14 say quarantine period starts from the moment you show symptoms or is the 7 days incubation period included in 14 day quarantine period?**

**Doctor:** after symptoms show. as per the nicd guidelines for clinical management of suspected or confirmed covid-19 disease, "those with mild disease may be de-isolated 14 days after symptom onset, while those with severe disease may be de-isolated 14 days after achieving clinical stability (e.g. once supplemental oxygen is discontinued)

Table 11: Examples of few-shot setting for English dataset.

患者描述：几天前左下腹剧痛，查出左输尿管结石。
患者：尿结石多长时间能排出来

医生： 结石的排出是没有具体的时间的，个人的情况不同排出的时间也会有差别。指导意见：你的情况是需要坚持治疗的，时间长没有排出可能由于结石比较大的缘故。相信大家都知道泌尿系结石对男人的伤害比较大，因此要尽快接受治疗，生活中要注意饮食问题，不要食用刺激食物，合理饮食。而且必要时可以定期复诊。

患者描述：腰椎椎体排列整齐，曲度变直，各椎体形态、信号未见异常；各椎间盘于T2WI信号未见减低，腰4-5椎间盘
患者：腰椎椎体排列整齐，曲度变直，各椎体形态

医生： 结石的排出是没有具体的时间的，个人的情况不同排出的时间也会有差别。指导意见：你的情况是需要坚持治疗的，时间长没有排出可能由于结石比较大的缘故。，相信大家都知道泌尿系结石对男人的伤害比较大，因此要尽快接受治疗，生活中要注意饮食问题，不要食用刺激食物，合理饮食。而且必要时可以定期复诊。

患者描述：致使小儿癫痫病的因素有哪些呢?小孩子怎么会得癫痫病呢，有时候经常睡午觉的时候半夜里惊醒，就会哭啼，不知晓怎么办 在乎怎样的帮助：致使小儿癫痫病的因素有哪些?
患者：引起孩子癫痫病的因素有哪些

医生： 一、脑危害与脑损伤在胚胎生长发育中受到病毒感染放射线映照，或其它缘由致使的胚胎发育不良能致使癫痫，胎儿生产过程中产伤也是致使癫痫的个主要缘由，颅脑外伤也可致使癫痫。二、隔代遗传要素是构成小儿癫痫的缘由地点，在许多有癫痫病史或有先天性中枢神经系统或心脏变形的患者宗族中简单呈现出癫痫。

患者描述：37了，之前查出脑发育不良，从小时候到现在身体状况良好，无抽搐，智力无影响，请问这疾病是保持现状还是会继续发展，对寿命有没有影响。
患者：脑发育不良会对寿命有影响吗

医生： 您好 现在治疗不晚的 一般3岁之内是最佳治疗时机越早越好的建议积极治疗 康复治疗配合营养脑细胞治疗 至于时间不好说 因为不知孩子的情况咋样 积极配合治疗 ，脑发育这种疾病不易痊愈。患者朋友应当保持良好的心态，用积极的心态去面对它，只有这样才能提高患者对抗脑发育的信心，相信这样一定能得到康复。

Table 12: Examples of few-shot setting for Chinese dataset.