BP4ER: Bootstrap Prompting for Explicit Reasoning in Medical Dialogue Generation

Yuhong He[1,6], Yongqi Zhang[2], Shizhu He[3,4] and Jun Wan[1,3,5,*]
[1] Macau University of Science and Technology, Macao, China
[2] 4Paradigm Inc., Beijing, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[4] The Lab of Cognition and Decision Intelligence for Complex Systems, CASIA, China
[5] MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[6] Zhongkai University of Agriculture and Engineering, Guangzhou, China
yuhonghe.ai@gmail.com, yzhangee@connect.ust.hk, shizhu.he@nlpr.ia.ac.cn, jun.wan@ia.ac.cn

Abstract

Medical dialogue generation (MDG) has gained increasing attention due to its substantial practical value. Previous works typically employ a sequence-to-sequence framework to generate medical responses by modeling dialogue context as sequential text with annotated medical entities. While these methods have been successful in generating fluent responses, they fail to provide process explanations of reasoning and require extensive entity annotation. To address these limitations, we propose the method **B**ootstrap **P**rompting for **E**xplicit **R**easoning in MDG (**BP4ER**), which explicitly model MDG's multi-step reasoning process and iteratively enhance this reasoning process. We employ a least-to-most prompting strategy to guide a large language model (LLM) in explicit reasoning, breaking down MDG into simpler sub-questions. These sub-questions build on answers from previous ones. Additionally, we also introduce two distinct bootstrapping techniques for prompting, which autonomously correct errors and facilitate the LLM's explicit reasoning. This approach eliminates the need for entity annotation and increases the transparency of the MDG process by explicitly generating the intermediate reasoning chain. The experimental findings on the two public datasets indicate that BP4ER outperforms state-of-the-art methods in terms of both objective and subjective evaluation metrics.

Keywords: Bootstrap Prompting, Medical Dialogue Generation, Explicit Reasoning

1. Introduction

Medical dialogue systems (MDS) are receiving significant attention due to the rising demand for telemedicine (Zhou et al., 2021; henfeng He et al., 2022), offering accessible medical services such as health consultations, diagnosis, and prescriptions, to a broader population (Yan et al., 2022; Xia et al., 2022b). Within MDS, medical dialogue generation (MDG) plays a crucial role by generating accurate medical responses based on given dialogue histories (Lin et al., 2021; Wei et al., 2018; Xu et al., 2019a). Typically, MDG involves understanding the patient's overall state, making the next diagnosis decisions in a limited-turn dialogue, and conducting medical reasoning analysis to generate responses (Li et al., 2021a; Chen et al., 2022).

Previous research on MDG typically adopts a framework in which dialogue context is modeled as sequential text (Xu et al., 2023; Liu et al., 2021), and medical entities are identified and annotated within this textual context (Liu et al., 2022b; Du et al., 2019b). Subsequently, response generation is carried out using sequence-to-sequence (Seq2Seq) models (Sutskever et al., 2014). These Seq2Seq methods leverage pre-trained text encoders and decoders to generate medical responses (Li et al., 2021a; Zhao et al., 2022), as illustrated in Figure 1 (a). Although these methods have yielded substantial success in generating coherent and fluent responses in MDG, they face two key challenges: (1) *Lack of process explanation.* To help patients or physicians understand why an MDG module generates a response, interpretability of the medical reasoning process is indispensable (Li et al., 2021a), i.e., information on patient status and diagnostic decision-making by physicians. (2) *Requirement for large-scale annotations.* Previous works (Xu et al., 2023; Zhao et al., 2022) heavily depend on the availability of a substantial amount of manually labeled data during the training phase. However, obtaining such data is often challenging due to the specialized medical knowledge required and stringent privacy considerations.

To address the limitations above, we propose the **B**ootstrap **P**rompting for **E**xplicit **R**easoning method (**BP4ER**), as illustrated in Figure 1 (b). Our motivation is to eliminate the need for entity annotation by treating MDG as a multi-step reasoning problem. Specifically, we explicitly break down MDG into a reasoning chain and sequentially address each intermediate reasoning step, aligning with its inherent multi-step reasoning process. Drawing from the concept of chain-of-thought prompting (Wei et al., 2022c), we introduce the least-to-most

* Corresponding author

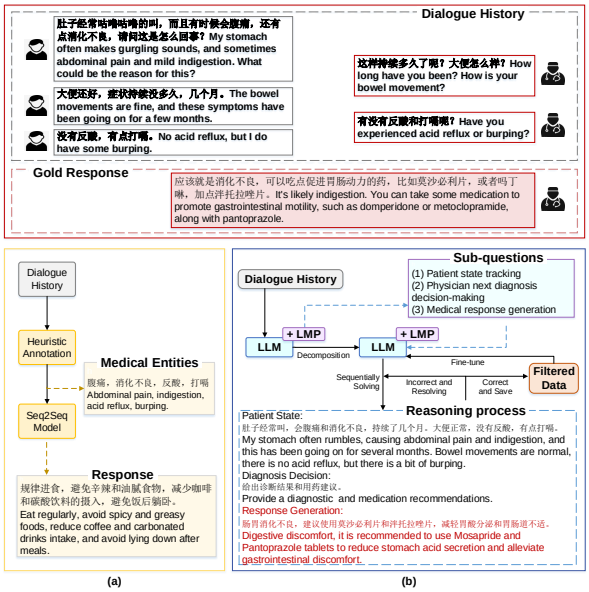

Figure 1: Paradigm comparison in MDG: prior works adopt a Seq2Seq framework (a); our model (b) explicitly incorporates a multi-step reasoning process and reduces entity annotation.

prompting (LMP) strategy (Zhou et al., 2023) to guide a large language model (LLM) (Zhao et al., 2023; Du et al., 2022) towards explicit reasoning in MDG. We first decompose the MDG process into a reasoning chain, comprising a series of interrelated sub-questions. Then, we follow Zelikman et al. (2022) and construct demonstration prompts for each sub-question and address them sequentially with answers from resolved sub-questions, promoting a coherent reasoning process.

Despite LLMs' impressive language understanding ability in general language modeling (Wang et al., 2022b; Huang and Chang, 2022; Zhu et al., 2023), their intermediate reasoning steps in MDG would be error-prone, reducing overall performance (Zhang et al., 2022). To facilitate the model's explicit reasoning ability, we propose two distinct bootstrapping techniques for prompting: answer-providing bootstrapping (AP-Bootstrap) and prompt-revising bootstrapping (PR-Bootstrap). These techniques allow the model to autonomously rectify errors without relying on large-scale annotations. Subsequently, we collect the accurate reasoning chain to create filtered data by implementing feedback loops. The model is then fine-tuned using this filtered data, and the process is repeated. This approach yields a significant improvement in the model's performance and enhances the quality of the generated responses.

Our contributions can be summarized as follows:

- We present a novel explicit reasoning model for medical dialogue generation (MDG) called BP4ER. To the best of our knowledge, BP4ER is the first model to systematically deconstruct MDG into an intermediate reasoning chain, which notably enhances the interpretability of the MDG process.
- BP4ER introduces the least-to-most prompting strategy to guide LLM for explicit reasoning and an iterative approach to bootstrap the prompting process for augmenting the LLM's reasoning capabilities, resulting in coherent and precise medical dialogue responses.
- We evaluate BP4ER on two public datasets using both automatic and manual evaluation metrics. Experimental results demonstrate its superiority over previous methods.

2. Related Work

2.1. Medical Dialogue Generation

Medical dialogue generation (MDG) has attracted increasing attention due to its high practical value. Early attempts at MDG were based on pre-defined templates to generate natural language (Ferguson et al., 2009; Wong et al., 2011; Xu et al., 2019b). However, template-based MDG suffers from the problem of inflexibility. Recently, Zeng et al. (2020) took an initial step in neural-based MDG. They pre-trained several dialogue generation models on large-scale medical corpora. Liu et al. (2022b) frame medical dialogue generation as entity prediction and entity-aware response generation. Furthermore, Liu et al. (2021) unifies the dialogue context understanding and entity reasoning through a heterogeneous graph. Li et al. (2021a) consider medical entities in the utterances as states and actions and present semi-supervised variation reasoning with a patient state tracker and a physician action network. Zhao et al. (2022) exploit the medical relationship between dialogue context and recall pivotal information to produce responses. Xu et al. (2023) models a medical entity flow and a dialogue act flow to improve entity selection and dialogue act prediction.

Although these models achieve comparable performance, they often lack process interpretability and need substantial annotation.

2.2. Prompt Learning of LLMs

Recent studies (Dong et al., 2023; Jeblick et al., 2022) have proposed various prompting strategies to strengthen and generalize the in-context learning ability of LLMs. One such strategy is chain-of-thoughts (CoT) prompting, introduced by Wei

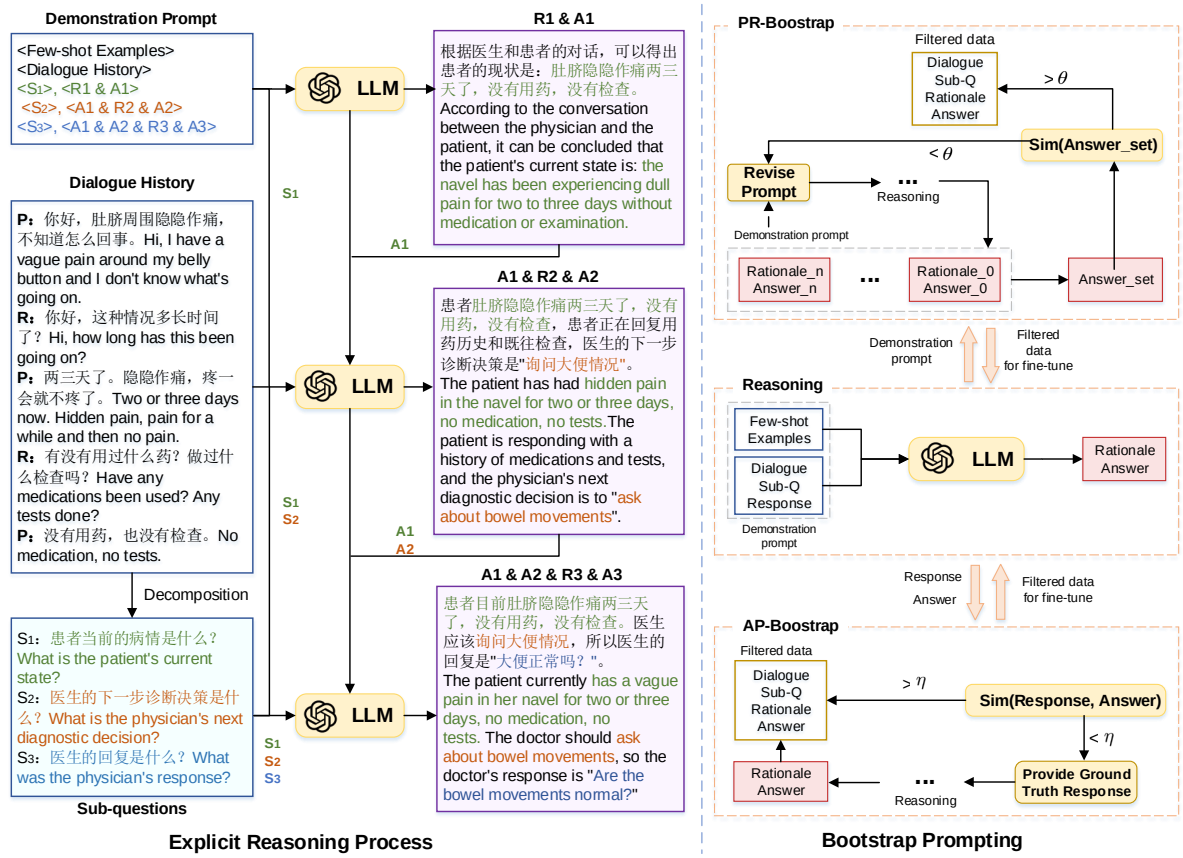

Figure 2: Overview of BP4ER. Medical dialogue is deconstructed into a reasoning chain of sub-questions. Demonstration prompts guide intermediate reasoning, sequentially querying the LLM. Two bootstrapping techniques for prompting, AP-Bootstrap and RP-Bootstrap, are introduced to enhance explicit reasoning.

et al. (2022c), which incorporates intermediate reasoning steps into LLMs to construct demonstrations between inputs and outputs. While Wei et al. (2022c) manually constructs CoTs, AutoCoT (Zhang et al., 2022) utilizes LLMs to automatically generate CoTs, using the prompt sentence "let's think step by step." Additionally, Wang et al. (2022a) propose iCAP, a context-aware prompter capable of dynamically adjusting contexts for each reasoning step. To tackle the challenge of easy-to-hard generalization, Zhou et al. (2023) propose a least-to-most prompting (LMP) strategy. Unlike CoT, which focuses on individual instances, LMP is task-oriented, breaking down a problem into interrelated sub-questions from a task perspective and forming a progressive prompt sequence for LLMs. Moreover, while CoTs are crucial for model performance, they are not readily available for specific tasks, and creating them requires significant time and resources, potentially introducing bias.

Inspired by LMP, we introduce MDG as a multi-step reasoning problem aimed at explicitly and iteratively modeling the reasoning process, mirroring the decision-making process of doctors in real medical scenarios.

3. Main Method

Problem Formulation. In the context of a dialogue comprising T turns, a medical dialogue session D is a sequence of utterances, denoted as $D = \{P_1, R_1, P_2, R_2, ..., P_T, R_T\}$. Here, P_t and R_t $(t = 1 \ldots T)$ refer to utterances from a patient and responses from a virtual physician, respectively. At the t-th turn, given the dialogue history $H = \{P_1, R_1, ...R_{t-1}, P_t\}$ as input, the model aims to generate an intermediate reasoning chain $S = \{S_1, ..., S_k\}$ and corresponding answers $A = \{A_1, ..., A_k\}$, where k is the number of reasoning steps. Subsequently, the model generates a medical response R_t for the current turn. Figure 2 provides an illustrative overview of our proposed BP4ER method. In this section, we provide a description of the multi-step reasoning process for MDG, as outlined in Section 3.1. Then, we present the details of the explicit reasoning process in Section 3.2, with a specific focus on augmenting the model's interpretability. Finally, we introduce two distinct bootstrapping techniques for prompting to enhance explicit reasoning in the BP4ER model, as discussed in Section 3.3.

3.1. Multi-step Reasoning

In real-world medical scenarios, MDG involves a multi-step reasoning process that aligns with the logical framework of medical consultation (Chen et al., 2022). It consists of three essential steps (Li et al., 2021a): (i) *Patient State Tracking:* Initially, the MDG system interacts with the patient to acquire additional symptoms beyond those self-reported. Here, the system focuses on comprehensively tracking and maintaining the patient's condition within the dialogue context, including symptoms, medications, and other relevant information. (ii) *Next Diagnosis Decision-making:* Drawing from the collected patient states and the ongoing conversation, the system infers the next diagnosis decision that a physician would make. This step guides the responses generated by the system, ensuring a coherent flow in the medical dialogue. (iii) *Medical Response Generation:* Utilizing the identified patient states and the diagnosis decision-making, the MDG system generates a contextually relevant and coherent response that aligns with the ongoing medical dialogue.

3.2. Explicit Reasoning Process

In Section 1, we emphasized the importance of explicitly demonstrating the multi-step reasoning process of MDG for better interpretability, rather than simply generating direct answers. To achieve this, we employ a few-shot Least-to-Most Prompting (LMP) strategy (Zhou et al., 2023) to guide the Large Language Model (LLM) (Zhao et al., 2023; Du et al., 2022). This strategy breaks down the complex MDG task into a sequence of interrelated sub-questions, inspired by medical diagnostic logic. In this study, we simplify this decomposition into three specific sub-questions following the multi-step reasoning process described in Section 3.1, creating an intermediate reasoning chain, denoted as $S = \{S_1, S_2, S_3\}$:

- S_1: What's the patient's current state?
- S_2: What's the physician's next decision?
- S_3: What's the physician's response?

As depicted in Figure 2, the process of generating a response from a dialogue history is reframed as answering two intermediate sub-questions: "*What's the patient's current state?*" and "*What's the physician's next diagnostic decision?*".

We tackle these sub-questions sequentially, with each solution building upon previously obtained answers. To facilitate this, we create question-rational-answer pairs as demonstrations and construct a demonstration prompt for each intermediate reasoning step, inspired by (Zelikman et al., 2022). This prompt consists of examples illustrating sub-question resolution, the dialogue history, a list of previously answered sub-questions and their corresponding answers (if any), and the next sub-question to be addressed.

The solving process starts with a few-shot prompting, providing the LLM with a demonstration prompt comprising few-shot examples, dialogue history, and the first sub-question. For example in Figure 2, the demonstration prompt is "*Examples: <Few-shot Examples>, Dialogue History H: P: Hi, I have a vague pain ... P: No medication, no tests, Sub-question S_1: What's the patient's current state?*". Then, We use the generated answer, e.g., "A_1: *The navel ... examination*," to construct the next prompt by appending the answer to the previous prompt followed by the next sub-question S_2: "*What's the physician's next diagnostic decision?*". This process repeats for sub-question S_3: "*What's the physician's response?*". The final answer (e.g., "A_3: *Are the bowel movements normal?*") for MDG R_t is obtained by adding the generated answer A_2 to the previous prompt. This approach allows us to address each sub-question sequentially, leveraging answers from previously resolved sub-questions, resulting in a coherent, step-by-step reasoning process.

3.3. Bootstrap Prompting

The intermediate reasoning steps in LLMs may contain errors, affecting reasoning results and overall performance. To enhance the explicit reasoning abilities of LLMs, drawing inspiration from (Wang et al., 2022a), we improve the quality of demonstrations through iterative prompting bootstrapping. During the training phase, the final step of reasoning benefits from having access to ground truth responses, ensuring accuracy. However, intermediate steps lack correct answers, posing a challenge. To overcome this limitation, we introduce two iterative bootstrapping techniques for prompting: answer-providing bootstrapping (AP-Bootstrap) and prompt-revising bootstrapping (PR-Bootstrap), tailored to different scenarios. AP-Bootstrap can be seen as a greedy decoding process, whereas PR-Bootstrap is based on a sampling approach. These techniques help LLMs to autonomously rectify errors in demonstrations, reducing the reliance on extensive annotations.

3.3.1. Answer-Providing Bootstrapping

Given a pre-trained LLM $\mathcal{M}$ and a dataset of dialogue histories $\mathcal{H}$ paired with responses $\mathcal{R}$, denoted as $\mathcal{D} = \{(H_i, R_i)\}_{i=1}^{N_D}$, the AP-Bootstrap approach takes a demonstration prompt as input. This prompt consists of a small example set $\mathcal{P}$, defined as $\mathcal{P} = \{H_i^p, Q_i^p, R_i^p\}_{i=1}^{N_P}$, where $N_P \ll N_D$ (e.g. $N_P = 5$).

Similar to standard few-shot prompting, this example set is concatenated with each dialogue history instance in $\mathcal{D}$ and sub-question S_i, resulting in $\hat{H}_i = \{H_1^p, Q_1^p, R_1^p, ..., H_{N_P}^p, Q_{N_P}^p, R_{N_P}^p, H_i, S_i\}$. This encourages the model to generate a rationale $\hat{Q}_i$ for H_i followed by an answer A_i. If the generated answers A_i are semantically similar to the gold response R_i, the reasoning process is considered credible. Otherwise, our objective is to correct the reasoning process and obtain available answers A_i. Finally, the credible and corrected dialogue data are combined for iterative fine-tuning of the LLM, enhancing its reasoning capabilities.

To achieve this, we employ cosine similarity, denoted as $Sim(.)$, to measure the semantic similarity between the generated answers A_i and the gold response R_i. We utilize this similarity metric to filter the dialogue data, retaining instances with high semantic similarity, i.e., $Sim(A_i, R_i) > \eta$, where η is a predefined threshold. For those instances with low similarity, following (Zelikman et al., 2022), we provide a model with the gold response, allowing it to autonomously rectify errors by generating a reasoning chain similar to the previous explicit reasoning process (as described in Section 3.2). By providing the gold response, the model can reason backward, facilitating the generation of a reasoning chain leading to the correct answer. After error correction, the dialogue with the revised reasoning chain is added to the filtered dataset. Subsequently, we fine-tune the LLM $\mathcal{M}$ on this filtered dataset and iteratively bootstrap prompting $\mathcal{M}$ to generate a new reasoning chain with the newly fine-tuned model until performance reaches a plateau. Throughout this iterative process, we consistently fine-tune from the original pre-trained model $\mathcal{M}$ to mitigate overfitting concerns.

The AP-Bootstrap method can be conceptualized as an approximation to an RL-style policy gradient objective. To illustrate this, consider that $\mathcal{M}$ can be interpreted as a discrete latent variable model $p_M(R|H) = \sum_Q p(Q|H)p(R|H,Q)$; in other words, $\mathcal{M}$ first samples a latent rationale Q before generating the response R. Now, given the indicator reward function $f_I = \mathbb{I}(Sim(A,R) > \eta)$, the total expected reward across the dataset is:

$$\mathcal{J}(\mathcal{M}, H, R) = \sum\nolimits_i \mathbb{E}_{\hat{Q}_i, A_i \sim p_M(\cdot|H_i)} f_I(\cdot)$$

whose gradient is obtained via the standard log-derivative trick for policy gradients:

$$\nabla \mathcal{J}(\mathcal{M}, H, R) = \sum\nolimits_i \mathbb{E}_{\hat{Q}_i, A_i \sim p_M(\cdot|H_i)} [f_I(\cdot) \cdot \nabla log p_M = (A_i, \hat{Q}_i|H_i)]$$

Note that the indicator function discards the gradient for dissimilar sampled demonstrations to the correct response R_i. Thus, the AP-Bootstrap approximates $\mathcal{J}$ by 1) greedily decoding samples of $(\hat{Q}_i, A_i)$ to reduce the variance of this estimate, and 2) taking multiple gradient steps on the same data batch, similar to policy gradient algorithms (Schulman et al., 2017).

3.3.2. Prompt-Revising Bootstrapping

During our experiments, we noticed that autonomous error correction faces challenges when dealing with complex dialogues, such as doctors continuously questioning patients, cross-questions between doctors and patients, and ambiguous descriptions of patient conditions. We attribute this challenge to the lack of correct answers in the intermediate steps of the reasoning process within MDG. To address this, we introduce a straightforward yet effective strategy called prompt-revising bootstrapping (PR-Bootstrap). This strategy capitalizes on the understanding that complex reasoning tasks often offer multiple pathways to arrive at a correct answer, as discussed in (Stanovich and West, 2000). In contrast to AP-Bootstrap, PR-Bootstrap alleviates the problem of limited diversity inherent in greedy decoding, as demonstrated in our experiments.

To implement PR-Bootstrap, we first prompt the LLM in the format of a demonstration prompt to yield an initial answer, which is added to the candidate answers. We then revise the few-shot examples in the original demonstration prompt to generate an alternative rationale, along with its corresponding new answer, which is also included in the candidate answers. It's important to note that each answer within the candidate set is derived from a distinct rationale. Therefore, if two answers exhibit significant semantic similarity, they are considered a consistent answer pair. We measure this similarity using cosine similarity calculations between the newly generated answer and those in the candidate set. Answer pairs surpassing a predefined threshold θ are considered the most consistent within the candidate answer set and are added to the filtered dataset. When no answer pairs meet the threshold θ, we iterate the prompt revision process to explore diverse reasoning paths and generate alternative answers until reliable answers are obtained for all provided data.

The iterative bootstrapping approach mirrors the human experience, where multiple different reasoning paths leading to the same answer increase confidence in its correctness. Finally, similar to the AP-Bootstrap method, we fine-tune the LLM on the filtered dataset to enhance its reasoning abilities by bootstrapping the prompting process.

4. Experiments

4.1. Datasets

We adopt two publicly available benchmark datasets, namely MedDG (Liu et al., 2022c) and KaMed (Li et al., 2021b), collected from medical consultation websites[1] after anonymization. **MedDG** contains 17K dialogues, focusing on 12 distinct diseases within the gastroenterology department. On average, each dialogue consists of 9.92 rounds. We divide the dataset into training/validation/test sets with sizes of 14,864/2,000/1,000 dialogues, as originally outlined in Liu et al. (2022b). **KaMed** contains over 63K dialogues, covering an extensive range of over 300 diseases across 13 different medical departments. KaMed exhibits a higher average dialogue length compared to MedDG, e.g., 11.62 rounds per dialogue. Following the setting in Xu et al. (2023), we filtered dialogues with privacy concerns and obtained 29,159/1,532/1,539 dialogues for the training/validation/test sets. The dataset presents challenging and diverse scenarios, with over 300 hospital departments.

4.2. Evaluation metrics

Automatic Evaluation. To evaluate the linguistic quality of the generated responses, we employ standard word-overlap-based metrics: BLEU (B@n) (Papineni et al., 2002) and ROUGE (R@n) (Lin, 2004). These metrics measure lexical quality by calculating n-gram overlaps between the generated and accurate responses. Additionally, we incorporate the DISTINCT (D@n) metric (Li et al., 2016) for a more comprehensive evaluation. DISTINCT-n measures response diversity by calculating the proportion of distinct n-grams within the generated responses, offering a valuable perspective on response quality often missed by traditional BLEU and ROUGE metrics.

Human Evaluation. Aligned with prior studies (Li et al., 2021a; Zhao et al., 2022), we conducted a human evaluation to assess the quality of responses in terms of fluency, coherence, and correctness. Fluency evaluation measures overall smoothness and naturalness, coherence assesses logical consistency with the dialogue history, and correctness measures the accuracy of medical knowledge in the responses. Consistent with Li et al. (2021a) and Zhao et al. (2022), we randomly sampled 100 cases and invited three professional annotators from a thirty-party hospital to perform manual evaluations. Annotators utilized the aforementioned metrics, rating each response on a scale from 1 (poor) to 5 (excellent). It's noteworthy that model names were anonymized to ensure objectivity throughout the evaluation process.

[1] https://www.chunyuyisheng.com/

4.3. Implementation Details

In this work, we used ChatGLM-6B[2] (Du et al., 2022) as the foundational LLM for BP4ER. ChatGLM-6B is equipped with 6 billion parameters and is optimized with the Adam optimizer (Kingma and Ba, 2014). We chose this model for its robust language understanding abilities in Chinese and its relatively lightweight design compared to other LLMs. Hyperparameters were selected based on the best-performing checkpoints during validation, with a batch size of 32 and a learning rate of 1e-2. For MedDG, we set similarity thresholds as [0.75, 0.8, 0.65] for its three reasoning steps, while for KaMed, they were [0.65, 0.75, 0.65]. All experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.4. Baseline models

Our method is compared with the following baselines. **Seq2Seq** (Sutskever et al., 2014) is an RNN-based sequence-to-sequence model with an attention mechanism. **HRED** (Serban et al., 2016) uses hierarchical encoders to model the dialogue context from token level and utterance level compared to Seq2Seq. **DialoGPT** (Zhang et al., 2019) and **GPT-2** (Radford et al., 2019) are transformers-based pre-trained language models widely adopted in tasks of dialogue generation. **VRBot** (Li et al., 2021a) summarizes patient states and physician actions into phrases through variational methods and generate the response. **MedPIR** (Zhao et al., 2022) exploit the medical relationship between dialogue context and recall pivotal information to produce responses in the recall-enhanced generator. **DFMed** (Xu et al., 2023) models the transitions of medical entities and dialogue acts with the pre-trained model. **ChatGLM-6B** (Du et al., 2022) is a pre-trained language model with 6 billion parameters, which generates medical responses directly.

4.5. Automatic Evaluation

Table 1 presents the automatic evaluation results for the MedDG and KaMed datasets, revealing several key insights:

(1) BP4ER demonstrates significant improvements, as depicted in Table 1. These results confirm BP4ER's efficacy in enhancing response quality and ensuring greater semantic consistency with

[2] It can be done with any off-the-shelf LLMs, such as LLaMA (Touvron et al., 2023) and Alpaca (Taori et al., 2023).

Dataset	Model	B@1	B@2	B@4	R@1	R@2	D@2
MedDG	Seq2Seq (Sutskever et al., 2014)	28.55	22.85	15.45	25.61	11.24	/
	HRED (Serban et al., 2016)	31.61	25.22	17.05	24.17	9.79	/
	DialoGPT (Zhang et al., 2019)	32.77†	26.93†	17.96†	27.11†	11.34†	79.26†
	GPT-2 (Radford et al., 2019)	35.27	28.19	19.16	28.74	13.61	/
	VRBot (Li et al., 2021a)	29.69	23.9	16.34	24.69	11.23	/
	MedPIR (Zhao et al., 2022)	38.72†	27.64†	18.14†	25.72†	10.30†	82.77†
	DFMed (Xu et al., 2023)	42.56	33.34	22.53	29.31	14.21	/
	ChatGLM-6B (Du et al., 2022)	37.96	24.22	15.37	18.05	10.53	89.81
	BP4ER (ours)	**44.78**	**33.80**	**23.76**	**41.47**	**22.47**	**89.93**
	Improvement	+2.22	+0.46	+1.23	+12.16	+8.26	+0.12
KaMed	Seq2Seq (Sutskever et al., 2014)	23.52	18.56	12.13	23.56	8.67	/
	HRED (Serban et al., 2016)	26.75	21.08	16.36	18.71	7.28	/
	DialoGPT (Zhang et al., 2019)	30.17†	25.53†	17.09†	24.30†	9.79†	80.27†
	GPT-2 (Radford et al., 2019)	33.76	26.58	17.82	26.8	10.59	/
	VRBot (Li et al., 2021a)	30.04	23.76	16.36	18.71	7.28	/
	MedPIR (Zhao et al., 2022)	29.42†	21.60†	16.47†	20.69†	9.27†	83.75†
	DFMed (Xu et al., 2023)	40.20	30.97	20.76	28.28	11.54	/
	ChatGLM-6B (Du et al., 2022)	38.70	27.19	16.38	33.86	20.21	85.70
	BP4ER (ours)	**41.89**	**31.74**	**20.81**	**35.76**	**21.19**	**86.83**
	Improvement	+1.69	+0.77	+0.05	+1.90	+0.98	+1.07

Table 1: Automatic evaluation (%) on MedDG and KaMed datasets. B@n denotes BLEU-n, R@n denotes ROUGE-n and D@2 denotes DISTINCT-2. The best values are in boldface and the second best are underlined. Models marked with † were reproduced by us, while the others were copied from the original results in (Xu et al., 2023).

gold standard responses. While ChatGLM-6B initially exhibits lower BLEU and ROUGE scores compared to DFMed, integrating explicit reasoning and bootstrapping prompting techniques yields notable enhancements. Specifically, there's a remarkable increase of 23.42% in ROUGE-1 and 11.94% in ROUGE-2. This integration not only boosts performance metrics but also enhances the transparency of the multi-step reasoning process in MDG. It renders the reasoning steps more comprehensible and interpretable without the need for extensive annotations. As a result, the model's decision-making process becomes more transparent, facilitating a deeper understanding of the underlying logic behind the generated responses.

(2) The performance enhancement is more pronounced in MedDG compared to KaMed, as indicated in Table 1. This discrepancy can be attributed to the fact that MedDG is focused on a specific department, i.e., gastroenterology, and contains a relatively small number of diseases, only 12 diseases. In contrast, KaMed covers a more extensive range of over 300 diseases across 13 different medical departments. The diversity and complexity inherent in KaMed render it a more challenging dataset for BP4ER. Additionally, it's worth noting that KaMed involves a greater number of dialogue rounds compared to MedDG, suggesting that the necessity to consider larger contextual information contributes to the dialogue's complexity. In summary, this observation suggests that BP4ER demonstrates more effectiveness when confronted with smaller and more focused datasets like MedDG, and it may encounter greater challenges when confronted with larger and more diverse datasets featuring extended dialogue rounds, such as KaMed.

(3) In comparison to traditional seq2seq-based models, LLM-based models demonstrate superior performance in the ROUGE and DISTINCT-2 metrics, while seq2seq-based models perform well on the BLEU metric. For instance, BP4ER achieves substantial improvements of 12.16% and 8.26% in the MedDG dataset when considering the ROUGE metric. Furthermore, LLM-based models consistently outperform other models in both the MedDG and KaMed datasets according to the DISTINCT-2 metric. These findings highlight the strength of LLM-based models in generating responses that closely align with the gold standard responses in terms of recall and content coverage, as indicated by the ROUGE metric. Additionally, they demonstrate the ability to produce diverse and distinct responses, as indicated by the DISTINCT-2 metric. Conversely, other models may prioritize response quality based on precision and n-gram matching, as indicated by their performance in the BLEU metric. In summary, the results underscore the strengths of LLM-based models in generating high-quality responses that capture both the richness and diversity in MDG,

Fine-Tune	Exp. Rea.	AP-Boots.	PR-Boots.	MedDG				KaMed			
				B@1	R@1	D@1	D@2	B@1	R@1	D@1	D@2
✓	✓	✓	✓	44.78	41.47	91.20	89.93	41.89	35.76	89.10	86.83
✓	✓	✓		42.27	37.64	89.76	88.73	40.69	35.01	87.97	85.94
✓	✓			40.75	36.63	90.14	88.90	39.68	34.99	88.47	86.07
✓				39.41	27.38	88.54	89.81	39.13	33.97	87.34	85.83

Table 2: Ablation studies (%) are carried out on two datasets by individually removing modules PR-Bootstrap, AP-Bootstrap and explicit reasoning process.

Model	Fluency	Cohe.	Correct.
DialoGPT	3.11	2.56	2.89
MedPIR	3.34	3.07	3.23
BP4ER	4.00	3.50	3.52
Gold	4.32	4.17	4.41

Table 3: Human evaluation (%) results on KaMed. The maximum score for each indicator is 5.

making them particularly suitable for tasks requiring comprehensive and diverse outputs.

4.6. Ablation Study

To assess the impact of different modules in BP4ER, we conducted ablation studies on two datasets by individually removing modules PR-Bootstrap, AP-Bootstrap, and the explicit reasoning process, as outlined in Table 2.

Firstly, we analyzed the effects of PR-Bootstrap on performance. Comparing the results to BP4ER, we observed decreases in all metrics upon removing PR-Bootstrap. This suggests that instructing the model of its incorrectness by revising the prompt positively influences model performance. Secondly, when removing AP-Bootstrap from BP4ER, we notice a slight increase in the DISTINCT-1/2 metric. We hypothesize that this improvement may be attributed to the fact that AP-Bootstrap can be considered as a form of greedy decoding, which tends to generate repetitive or monotonous sequences rather than diverse content. Finally, upon removing the explicit reasoning process, we observed a decline in all evaluation metrics, with a notably significant drop of 8.3% in the ROUGE metric in the MedDG dataset. This indicates that the introduction of an explicit reasoning process enhances the interpretability of the response generation in the MDG and improves the semantic similarity between the generated response and the ground truth.

Our ablation experiments robustly confirm the effectiveness of each module on model performance. The results indicate that all these modules contribute positively to our approach, underscoring their importance in achieving superior performance in MDG tasks.

4.7. Human Evaluation

In addition to quantitative evaluations, we conducted a human evaluation to assess the responses generated by different models in terms of fluency, consistency, and entity correctness. We randomly sampled 100 instances from the test set of KaMed, and the corresponding responses were generated by well-performing models such as DialoGPT, MedPIR, and BP4ER. To ensure the fairness of the assessment, the responses for each sample were shuffled and then presented to volunteers for evaluation. The final statistical results are summarized in Table 3. Notably, our proposed model BP4ER consistently outperformed other models across all three manual evaluation indicators. Particularly noteworthy is BP4ER's superiority in fluency and coherence, suggesting that our proposed method significantly enhances the quality of responses. This improvement can be attributed to the explicit decomposition of MDG's multi-step reasoning process and the iterative bootstrapping on prompting, both of which contribute to generating more linguistically fluent and contextually coherent responses, improving the overall user experience in medical dialogue scenarios. These findings provide additional evidence of the effectiveness of BP4ER in improving the quality of responses in MDG tasks, demonstrating its superiority over existing models.

4.8. Case Study

We randomly selected dialogue examples from the KaMed test set and compared the generated responses from several models in Figure 3.

In case 1, BP4ER follows a multi-step reasoning procedure: firstly, it summarizes the patient's current state and then infers the next diagnostic action for the physician, i.e., inquire about the patient's medical history. Finally, BP4ER generates a response semantically close to the gold response, both emphasizing "asking for the specific area of pain". In contrast, other models primarily focus on providing suggestions. This highlights BP4ER's ability to produce responses that are more aligned with the context of the dialogue, achieved through explicit reasoning in MDG and iterative bootstrap prompting.

P: I have morning and evening stomach pain, which eases after eating, with normal bowel movements and no nausea.
R: How long? Have any prior stomach issues?
P: Three or four days. No stomach issues, just these two days.
Gold Response: Probably peptic ulcer. You can take Daxi and omeprazole .
MedPIR: Stomach indigestion; eat small, frequent meals.
ChatGLM: Gastric ulcer, avoid greasy, spicy foods and caffeine.
BP4ER:
Patient's State: Stomach pain in the morning and evening, eases after eating, normal stools, no nausea, lasts for three or four days, no stomach issues.
Diagnostic Decision: Give diagnostic results and drug advises.
Response: Gastric ulcer, suggest to take anti-inflammatory drugs, such as omeprazole.

P: Is the pain on the outer front of the knee rheumatism?
Gold Response: How long? Does it hurt when pressed down?
MePIR: The knee pain can have many causes, not necessarily related to rheumatism.
BP4ER: How long has the pain been lasting? How does it pain?

P: What medication or injection for brucellosis? I'm a sheep farmer, and I get the disease with no energy and cold.
R: Hi, have you been diagnosed? How long have you been sick?
P: It has been two or three months since the diagnosis.
Gold Response: Rifampicin 600-900mg/day plus tetracycline 200mg/day for 6 weeks. Is there any medication? It needs a sufficient course of treatment.
ChatGLM: Brucellosis is an infectious disease caused by Brucella, that may require antibiotics to treat.
BP4ER: Have any medicine? Antibiotics are usually used for brucellosis, like tetracycline.

Figure 3: A case study on comparative responses generated from various models, where "P" represents patient descriptions and "R" represents system responses.

In case 2, upon receiving the patient's self-reported information, BP4ER predicts the next diagnostic decision and generates a response to inquire about the patient's drug history. This rationale closely aligns with medical logic. Conversely, other models offer advice without a comprehensive understanding of the patient's medical background, lacking medical rationale.

In case 3, it is apparent that ChatGLM-6B is limited to providing only approximate antibiotic drug recommendations. Conversely, BP4ER exhibits a more advanced ability by not only inquiring about the patient's medication history but also providing specific recommendations for antibiotic drugs suitable for the individual's condition. Despite this advanced ability, BP4ER still falls short when compared to the gold standard response, particularly in accurately determining the appropriate dosage and duration for medication use. This finding underscores the crucial necessity of integrating expert medical knowledge into the model to achieve precision for effective medical decision-making.

5. Conclusion

In this paper, we propose BP4ER, a novel medical dialogue generation (MDG) model. BP4ER employs a least-to-most prompting strategy to guide a large language model (LLM) towards explicit reasoning. This strategy involves breaking down MDG into a sequence of interrelated sub-questions, making the process closer to real medical reasoning. Each sub-question is driven by answers obtained from resolving preceding queries. Furthermore, the model incorporates two iterative bootstrapping techniques for prompting, enhancing the LLM's explicit reasoning ability. Through the iterative approach, BP4ER autonomously corrects intermediate errors, leading to more precise and coherent medical responses. These features collectively enhance the transparency and interpretability of the medical reasoning process while improving the overall quality of the generated medical dialogue responses. Both automatic and human evaluations consistently show BP4ER outperforming existing state-of-the-art methods.

6. Limitations

Given that the BP4ER relies on large language models and prompts to direct response generation, it necessitates greater computational resources to execute the reasoning chain and bootstrap prompting prior to generating responses. Another crucial limitation lies in the potential for the model to generate incorrect or nonsensical responses during the reasoning process. This risk arises from the inherent reliance on the reasoning ability of LLMs, which possess general knowledge but lack the specialized medical knowledge for accurate medical dialogue generation. As a result, there's a notable gap between the model's understanding of medical concepts. In future work, we hope to explore the introduction of medical knowledge to further enhance the model's explicit reasoning ability in the medical domain.

7. Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2021YFE0205700, Beijing Natural Science Foundation JQ23016, the External cooperation key project of Chinese Academy Sciences 173211KYSB20200002, the Science and Technology Development Fund of Macau Project 0123/2022/A3, and 0070/2020/AMJ, the National Natural Science Foundation of China (No.62376270), Open Research Projects of Zhejiang Lab No. 2021KH0AB07 and CCF-Zhipu AI Large Model Project 202219.

8. Bibliographical References

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

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022. Mdiaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of AAAI*, pages 4432–4440.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019a. Extracting symptoms and their status from clinical conversationsextracting symptoms and their status from clinical conversations. In *Proceedings of ACL*, page 915–925.

Nan Du, Mingqiu Wang, Linh Tran, Gang Lee, and Izhak Shafran. 2019b. Learning to infer entities, properties and their relations from clinical conversations. In *Proceedings of EMNLP-IJCNLP*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of ACL*, pages 320–335.

George Ferguson, James F. Allen, Lucian Galescu, Jill Quinn, and Mary D. Swift. 2009. Cardiac: An intelligent conversational assistant for chronic heart failure patient heath monitoring. In *AAAI Fall Symposium: Virtual Healthcare Interaction*.

henfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chenand Guandong Xu, , and Jian Wu. 2022. Dialmed: A dataset for dialogue-based medication recommendation. In *Proceedings of COLING*, page 721–733.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Proceedings of NIPS*, 35:3843–3857.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021a. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of SIGIR*, pages 11–15.

Li, Dongdong and Ren, Zhaochun and Ren, Pengjie and Chen, Zhumin and Fan, Miao and Ma, Jun and de Rijke, Maarten. 2021b. *Semi-Supervised Variational Reasoning for Medical Dialogue Generation*. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. PID https://github.com/lddsdu/VRBot.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and B Dolan William. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL*, pages 110–119.

Chinyew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of AAAI*.

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 EMNLP-IJCNLP*, page 5033–5042.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Proceedings of NIPS*, 35:1950–1965.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022b. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Liu, Wenge and Tang, Jianheng and Cheng, Yi and Li, Wenjie and Zheng, Yefeng and Liang, Xiaodan. 2022c. *MedDG: an entity-centric medical consultation dataset for entity-aware medical dialogue generation*. CCF International Conference on Natural Language Processing and Chinese Computing. PID https://github.com/lwgkzl/MedDG.

Wenge Liu, Jianheng Tang, Xiaodan Liang, and Qingling Cai. 2021. Heterogeneous graph reasoning for knowledge-grounded medical dialogue system. In *Neurocomputing*, page 260–268.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022d. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of ACL*, pages 61–68.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, and David Luan. 2021. Show your work: Scratchpads for intermediate computation with language models. In *arXiv:2112.00114*.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 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 Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of COLING*, pages 311–318.

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.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *Proceedings of ICAST*, volume 4, page 1.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of ACL*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*, volume 30.

Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *Proceedings of AAAI*, page 8838–8845.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Keith E Stanovich and Richard F West. 2000. 24. individual differences in reasoning: Implications for the rationality debate? *Behavioural and Brain Science*, 23(5):665–726.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Proceedings of NIPS*, 27.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. *arXiv preprint arXiv:2203.08383*.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022b. What language model architecture and pretraining objective works best for zero-shot generalization? In *Proceedings of ICML*, pages 22964–22984. PMLR.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language model with self generated instructions. In *Proceedings of ACL*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of NIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022c. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of NIPS*, 35:24824–24837.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of ACL*, pages 201–207.

Wilson Wong, John Thangarajah, and Lin Padgham. 2011. Health conversational system based on contextual matching of community-driven question-answer pairs. In *Proceedings of CIKM*, pages 2577–2580.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022a. Lingyi: Medical conversational question answering system based on multi-modal knowledge graphs. *arXiv e-prints*, pages arXiv–2204.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022b. Medconqa: Medical conversational question answering system based on knowledge graphs. In *Proceedings of EMNLP*, pages 148–158.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of AAAI*, page 1062–1069.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. 2023. Medical dialogue generation via dual flow modeling. In *ACL*, pages 33–40.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019a. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of AAAI*, page 7346–7353.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019b. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of AAAI*, volume 33, pages 7346–7353.

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. In *Proceedings of SIGIR*, pages 3013–3024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proceedings of NIPS*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and

Pengtao Xie. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of EMNLP*, page 9241–9250.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, , Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of ACL*, page 6460–6469.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. 2022. Medical dialogue response generation with pivotal information recalling. In *Proceedings of KDD*, pages 14–18.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompt enables complex reasoning in large language models. In *Proceedings of ICLR*.

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, et al. 2021. On the generation of medical dialogs for covid-19. In *Proceedings of ACL-IJCNLP*.

Fangqi Zhu, Yongqi Zhang, Lei Chen, Bing Qin, and Ruifeng Xu. 2023. Learning to describe for predicting zero-shot drug-drug interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14855–14870.