Guiding Clinical Reasoning with Large Language Models via Knowledge Seeds

Jiageng Wu[1], Xian Wu[2], Jie Yang[3]
[1]Zhejiang University, Hangzhou, China
[2]Tencent Youtu Lab, Jarvis Research Center, Beijing, China
[3]Brigham and Women's Hospital, Harvard Medical School, Boston, USA
jiagengwu@zju.edu.cn, kevinxwu@tencent.com, jieynlp@gmail.com

Abstract

Clinical reasoning refers to the cognitive process that physicians employ in evaluating and managing patients. This process typically involves suggesting necessary examinations, diagnosing patients' diseases, and selecting appropriate therapies, etc. Accurate clinical reasoning requires extensive medical knowledge and rich clinical experience, setting a high bar for physicians. This is particularly challenging in developing countries due to the overwhelming number of patients and limited physician resources, contributing significantly to *global health inequity* and necessitating automated clinical reasoning approaches. Recently, the emergence of large language models (LLMs) such as ChatGPT and GPT-4 have demonstrated their potential in clinical reasoning. However, these LLMs are prone to hallucination problems, and the reasoning process of LLMs may not align with the clinical decision pathways of physicians. In this study, we introduce a novel framework, In-Context Padding (ICP), to enhance LLMs reasoning with medical knowledge. Specifically, we infer critical clinical reasoning elements (referred to as knowledge seeds) and use these as anchors to guide the generation process of LLMs. Experiments on two clinical question datasets validate that ICP significantly improves the clinical reasoning ability of LLMs.

1 Introduction

Clinical reasoning is a pivotal process where healthcare professionals incorporate clinical evidence and medical knowledge to assess, diagnose, and decide on treatment for patients [Montgomery Jr, 2018]. It entails a series of cognitive tasks, including gathering patient information, formulating and evaluating diagnostic hypotheses, and making treatment decisions [Young *et al.*, 2018]. Consequently, clinical reasoning requires extensive medical knowledge and rich clinical experience, setting high expectations for physicians. However, in low- and middle-income countries (LMICs), medical resources are often scarce, making it difficult to access high-quality medical care. Despite accounting for 90% of the global burden of disease, LMICs only contribute to 12% of global health spending [Gottret and Schieber, 2006]. Furthermore, in 98 countries across Asia and Africa, the population of physicians does not meet the minimum threshold required for achieving 80% universal health coverage, underscoring a critical shortage of qualified clinical specialists [Haakenstad *et al.*, 2022].

The shortage of healthcare resources urges the emergence of automated approaches with reliable clinical reasoning capabilities to support clinical decisions. Recently, Large Language Models (LLMs) have shown great potential in the medical domain [Lee *et al.*, 2023], such as medical education [Lee, 2023], online consultation [Wu *et al.*, 2024c], and clinical report summarization [Nayak *et al.*, 2023]. Encouragingly, several advanced LLMs qualified for the medical licensing examinations at high scores, such as Med-PaLM [Singhal *et al.*, 2023], ChatGPT [Wu *et al.*, 2024b], and GPT-4 [Nori *et al.*, 2023], indicating remarkable proficiency in medical knowledge and clinical case analysis.

Despite the notable capabilities in comprehending humans' intentions and generating coherent responses [Zhao *et al.*, 2023], directly applying LLMs to the medical field has also raised concerns over the generations of incorrect knowledge and hallucination during clinical reasoning [Bernstein *et al.*, 2023; Liu *et al.*, 2024]. This primarily stems from these advanced LLMs being predominantly trained on general-domain data [Wu *et al.*, 2024a]. Lacking extensive training in domain-specific text, they fail to encode sufficient medical expertise and comprehend medical texts laden with specialized concepts [Liévin *et al.*, 2022]. Absent a solid foundation of medical knowledge, LLMs struggle to grasp the intricate medical context and identify critical concerns behind it, making it difficult to generate comprehensive medical inferences.

To tackle these challenges, we propose a novel framework **I**n-**C**ontext **P**adding (**ICP**) to enhance the inference capacity of LLM in the context of clinical reasoning. ICP consists of four major steps: 1) ICP firstly extracts medical entities from the clinical context and the reasoning objective; 2) In cooperation with the knowledge graph (KG), ICP then infer relevant medical entities (referred to as knowledge seeds) which could be helpful in clinical reasoning; 3) The acquired knowledge seeds are padded to the prompt and used to guide the inference process of LLMs; 4) finally, LLMs generate the clinical reasoning results as well the detailed explanation of how this reasoning is conducted. Extensive experiments and analyses on two datasets highlight a significant improvement in both the accuracy and interpretability of LLM. The ICP framework incorporates KG and in-context learning of LLM to efficiently bridge the knowledge gap in medical scenarios, ensuring its broad applicability in specialized domains. In addition to clinical reasoning results, the proposed ICP also provides an explanation of the reasoning process. Overall, the contributions are as follows:

- We propose **I**n-**C**ontext **P**adding (ICP) which enhances LLMs to conduct clinical reasoning. This is especially beneficial for less developed countries where high-quality medical care is hard to access.
- We infer the knowledge seeds from context information which are used as anchors for LLMs to conduct reasoning. This helps to align the LLM generation with the clinical reasoning process of physicians.
- Experimental results on two datasets validate the effectiveness of the proposed ICP. In addition to the final answer, ICP also provides a description of the reasoning process, making it more transparent and understandable.

2 Related Work

2.1 LLMs in Medicine

As several advanced LLMs have passed the medical licensing examination, there is a growing interest among researchers to explore the deployment of LLMs in clinical environments. Bernstein *et al.*(2023) and Ayers *et al.*(2023) compared the responses of ChatGPT with medical experts to healthcare questions posed by patients, which indicated the LLM could potentially offer helpful suggestions across various patient inquiries. Moreover, the LLM's responses were also rated significantly higher in terms of both quality and empathy. For multi-modality tasks in medicine, ChatCAD [Wang *et al.*, 2023b] was developed to incorporate the LLMs for an interactive computer-aided diagnosis of medical images. Jeblick *et al.*(2022) and Lyu *et al.*(2023) assessed the capabilities of LLMs in translating radiology reports into plain language to be easier understood by patients.

However, the LLM may provide false knowledge and hallucinations due to the misleading brought by the training corpus [Bernstein *et al.*, 2023; Zhao *et al.*, 2023]. This makes them challenging to deploy in a clinical setting, which necessitates rich medical knowledge and experience to conduct rigorous reasoning. While several studies attempted to integrate knowledge to enhance performance in medical tasks [Wu *et al.*, 2024b; Gao *et al.*, 2023], they typically required high-quality, large-scale, and structured medical knowledge for retrieval, which in turn limits their application. Furthermore, many of them focus on accuracy, overlooking the desire to improve medical inference during problem-solving, a crucial aspect of integrating LLMs into clinical practice.

2.2 Reasoning with LLMs

Owing to the exceptional ability in text generation and in-context learning, Wei *et al.* (2022) elicited the LLM to generate a detailed reasoning process before offering answers to questions under a few-shot setting, called Chain-of-Thought (CoT). It enhanced both accuracy and interpretability in various reasoning tasks. Then, Wang *et al.*(2022)explored the potential of sampling multiple reasoning paths, combining them using an ensemble voting technique to bolster performance. Meanwhile, Kojima *et al.*(2022) introduced a simple prompt, ”*Let's think step by step*”, to prompt LLMs to elucidate their analysis and then conclude with the answer, without any manually crafted examples.

Addressing the challenges of complex multi-step reasoning, numerous novel frameworks were developed to enhance the logical reasoning ability of the LLMs. Yao *et al.*(2022) decomposed the entirety of task-solving into reasoning and acting steps to progressively complete it. Taking a different approach, the Tree of Thought (ToT) [Yao *et al.*, 2023] empowered LLMs to consider multiple different reasoning paths and self-evaluate choices to decide the next course of action, achieving notable performance improvements in tasks like the Game of 24 and word-based games. Furthermore, the Graph of Thought (GoT) [Besta *et al.*, 2023] formulates the reasoning process using a graph to enhance the problem-modeling capability of CoT. Nonetheless, these studies mainly concentrated on mathematical reasoning and word games. There is limited exploration in the medical domain, where problem-solving also requires rigorous reasoning.

3 Methodology

3.1 Problem Formulation

In this subsection, we formulate the clinical reasoning task. We denote an instance of clinical reasoning with four elements: $\{Q, O, R, A\}$, where Q refers to the context information and the reasoning objective, $O = \{o_0, ..., o_s\}$ refers to s candidate options, A refers to the correct answer which is one option in O, and R refers to the detailed analysis to conclude A. Given Q and options O, conducting clinical reasoning can be formulated as estimating the probability of generating reasoning steps and then determining the correct answer $P(R, A|Q, O)$. As shown in Figure 1, Q refers to 1) the conditions of this patient; 2) the reasoning objective which is to infer the most appropriate treatment; O refers to five candidate drugs for treatments: Artane, Levodopa, Selegiline, Amantadine and Bromocriptine; A refers to the best choice Levodopa; R refers to the description of reasoning process.

3.2 Overall Workflow

Figure 1 illustrates the overall workflow of ICP. The fundamental principle of ICP is to identify the most relevant knowledge seeds and incorporate them into the in-context prompt as potential anchors. These anchors guide the construction of a multi-step inference path. The process comprises four main steps: 1) Identification of medical entities to understand the context information and the primary object for reasoning; 2) Mining of the potential knowledge seeds by incorporating identified medical entities and a medical knowledge graph learned from historical cases; 3) Composition of the whole prompt for LLM, including task instruction, context, options, and mined knowledge seeds; 4) Guiding clinical reasoning

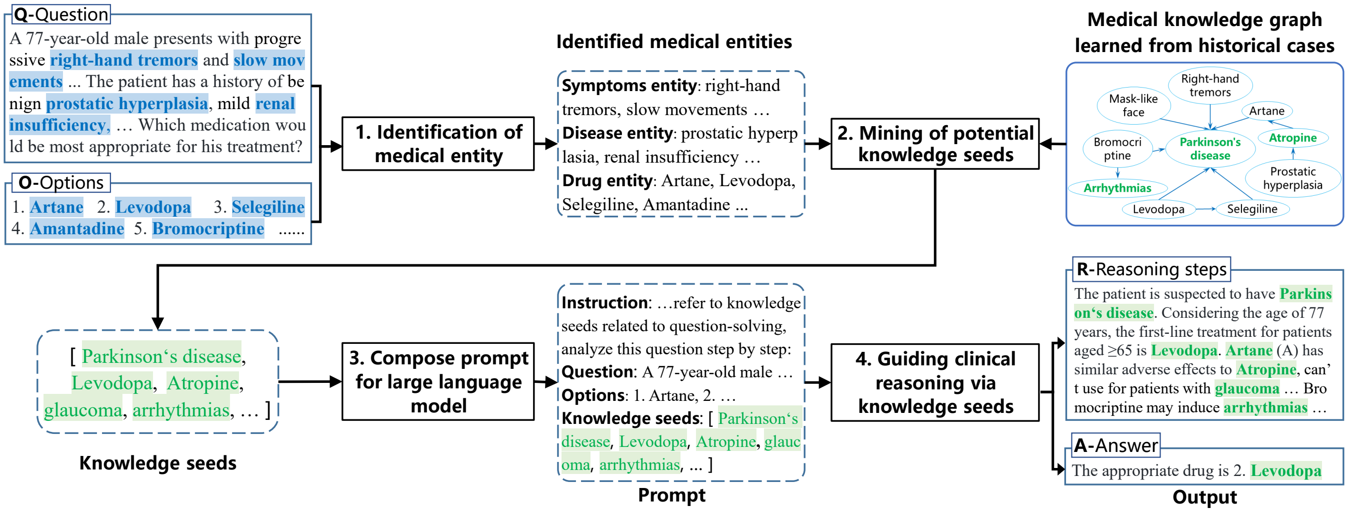

Figure 1: Overall Workflow of In-context Padding. It includes four main steps: 1) Identification of medical entities from a question and its candidate options; 2) Mine potential knowledge seeds by incorporating the identified medical entities and a medical knowledge graph learned from historical cases; 3) Compose the whole prompt for LLM, including task instruction, question, options, and mined knowledge seeds; 4) Guide LLMs to conduct clinical reasoning, leveraging knowledge seeds as anchors to perform relevant inferences and arrive at a conclusion. The blue text highlights key medical concepts in the question and options, while the green text indicates the identified knowledge seeds not in the question but critical for the reasoning process.

with LLM via knowledge seeds, which serve as anchors to conduct relevant inferences and conclude the final answer.

3.3 Identification of Medical Entities

For each instance, we concatenate Q and O and then extract the medical entities, including disease, symptoms, drugs, and any medical concepts, which briefly but precisely represent the medical context and the reasoning target. For the training instances, we also extract the medical entities discussed in its detailed analysis R.

To effectively extract the medical entities, we harness the exceptional in-context learning capabilities of LLMs and few-shot learning. Specifically, we provide the LLM with detailed instructions and representative examples to demonstrate the extraction process. This approach eliminates the time-consuming and labor-intensive need for manual annotation and specific model fine-tuning, making it easily adaptable to various clinical questions [Tarcar *et al.*, 2019; Agrawal *et al.*, 2022]. We select Baichuan2-7B-Chat [Yang *et al.*, 2023] to identify medical entities, which is an advanced open-source LLM and demonstrated excellent performance across multiple Chinese tasks. More details can be found in the code repository[1].

3.4 Construction of Medical Knowledge Graph

In this subsection, we construct a medical knowledge graph, $G(E, V)$, to encapsulate the relationship among various medical entities, where V and E denote the sets of m nodes $\{e_1, ..., e_m\}$ and edges v_{ij} of G, respectively. Each node in this graph represents a medical entity. The directed edge v_{ij} indicates the likelihood of discussing e_j within an analysis given the presence of e_i in the question and options. It is mathematically represented as:

$$v_{ij} = P(e_j \in R \mid e_i \in \{Q, O\}) \tag{1}$$

An effective approach to compute the value of edge v_{ij} is counting the instances of e_i in Q or O and e_j in the corresponding analysis R. Without any training, this process can be directly calculated based on a training dataset encompassing n instances:

$$\hat{v}_{ij} = \sum_{k=0}^{n} f(e_i, e_j, Q_k, O_k, R_k) \tag{2}$$

where $\{Q_k, O_k, R_k, A_k\}$ denotes a training sample k (refer to section 3.1). f is defined as:

$$f(e_i, e_j, Q_k, O_k, R_k) = \begin{cases} 1, & if\ e_i \in \{Q_k, O_k\}\ and\ e_j \in R_k \\ 0, & otherwise \end{cases} \tag{3}$$

Notably, due to the inherent directionality of the edges, v_{ij} and v_{ji} delineate the inverse path, leading to distinct values. If no relationship is established between node e_i and node e_j, the $\hat{v}_{ij}$ is designated a value of -1.

Some common entities have been frequently mentioned in many questions. To enhance the specificity between associated entities, we implement a weighting scheme. The resultant weighted edge value, v_{ij}, can be formulated by:

$$v_{ij} = \frac{\hat{v}_{ij}}{\sum_{k=1}^{m} \hat{v}_{ik}} * \lg[\frac{m}{1 + c_j}] \tag{4}$$

Here, c_j denotes the frequency with which entity e_j appears in the analysis of the training dataset:

$$c_j = \sum_{i=1}^{n} g(e_j, R_k) \tag{5}$$

[1] https://github.com/Dragon-Wu/ICP-for-Clinical-Reasoning

$$g(e_j, R_k) = \begin{cases} 1, & if\ e_j \in R_k \\ 0, & otherwise \end{cases} \tag{6}$$

To represent the entities and edges associated with e_i, we introduce the $V_{e_i} \in V$ and $E_{e_i} \in E$ to denote all edges sourced from e_i and their targeted nodes.

3.5 Mining Potential Knowledge Seeds

The essence of clinical context information can be approximately captured by the identified entities within Q and O. We extract the entities X from $\{Q, O\}$. For each extracted entity $x_i \in X$, the edges started from it V_{x_i} and the targeted nodes E_{x_i} are identified. Consequently, the aggregated sets of connected entities and edges are:

$$E_X = \{E_{e_1} \cup E_{e_2}, \cup ..., \cup E_{e_i}\} \tag{7}$$

$$V_X = \{V_{e_1} \cup V_{e_2}, \cup ..., \cup V_{e_i}\} \tag{8}$$

For each targeted entity $e \in E_X$ and its corresponding edge $v \in V_X$, the correlation of e to an entity x_i present in the question or options is computed. This relevance is determined by locating e in E_x and subsequently discerning the position of v in the sorted V_x:

$$q(e, v, E_x, V_x) = \begin{cases} rank(v, V_x), & if\ e \in E_x \\ inf, & otherwise \end{cases} \tag{9}$$

Subsequently, we accumulate the ranks of e concerning all entities in X to estimate the overall correlation of e to all entities mentioned in Q and O:

$$q(e, X) = \sum_{i=1} q(e, v, E_{x_i}, V_{x_i}) \tag{10}$$

Finally, we prioritize the entities $e \in E_X$ based on their value of $q(e, X)$ and pinpoint the top 10 entities as the *knowledge seeds*, which are most likely to be discussed in inferences.

3.6 Guiding Clinical Reasoning via Knowledge Seeds

After detecting the knowledge seeds, we incorporate them into the prompt to steer the LLM toward producing a reasonable analysis. Specifically, we append these identified knowledge seeds following the questions and options. We then instruct the LLM with the detailed description, "*Here is a clinical question, please refer to the knowledge seeds related to question-solving, and analyze this question step by step.*" This ensures that the LLM focuses and deliberates upon these entities, resulting in more concrete inference.

While we underscore the significance of these knowledge seeds, the LLM might encounter challenges in understanding or effectively using them. Moreover, certain entities may be solely related to the entities within Q and O but do not facilitate its solution. To efficiently guide the LLM in harnessing these knowledge seeds, we adopt a few-shot learning strategy. Representative samples in the format of {*question, option, knowledge seeds, analysis, and answer*} are provided. Consequently, these exemplars guide the LLM to recognize helpful seeds and progressively factor them into the inferences, yielding a reasonable analysis and correct answer.

Statistics	CNMLE-Clinical		CMExam	
	Train	Test	Train	Test
Number of questions	14655	600	56965	600
Average length of question and options	101.23	102.71	85.56	86.70
Average length of analysis	89.26	96.54	220.02	208.82
Number of entity within question and options	14.57	14.79	13.57	13.11
Number of entity within analysis	11.39	12.56	29.09	25.33

Table 1: Dataset Statistics.

4 Experiments and Results

4.1 Dataset

The Chinese National Medical Licensing Examination (CNMLE)[2] serves as the official certification examination for medical practitioners in China, akin to the United States Medical Licensing Examination (USMLE). It is a prerequisite for the candidates to have undergone five years of medical education, in addition to at least one year of assessed clinical practice. The objective of the CNMLE is to evaluate the practical abilities in a real-world clinical setting. Therefore, we built two clinical reasoning data sets from CNMLE. The dataset statistic is shown in Table 1.

CNMLE-Clinical

There are different examinations for different clinical disciplines, the first dataset *CNMLE-Clinical* focuses on CNMLE for clinical medicine. The CNMLE-Clinical evaluates four parts of medicine: preventive medicine, preclinical medicine, clinical medicine, and medical humanities, which cover over twenty distinct medical subjects. We gathered questions from past examinations and various reference books, accumulating a total of 15,255 questions. Out of these, we randomly selected 600 questions, consistent with the number in the official examination, to serve as our testing set, while the remainder were used as the training set. Each instance in CNMLE-Clinical consists of a question, five candidate options, the correct answer, and a detailed explanation for the answer.

CMExam

The second dataset, *CMExam* [Liu *et al.*, 2023], is a more comprehensive collection encompassing the six types of medical licensing examination: clinical medicine, traditional Chinese medicine (TCM), integrated TCM and Western medicine, dentistry, public health, pharmacy, and traditional Chinese pharmacy. The CMExam incorporates 68,119 medical questions sourced from past examinations and medical books. To ensure consistency with the genuine examination, we included 57,565 questions that have five options, a singular correct answer, and an analysis exceeding 30 words. Lastly, we randomly selected 600 questions to serve as the testing set, while the remaining questions in CMExam were designated as the training set. As shown in Table 1, the analyses of CMExam usually have more details than CNMLE-Clinical, which contains 25-29 entities in analyses, significantly higher than CNMLE-Clinical.

[2]https://www1.nmec.org.cn/Pages/ArticleInfo-13-10706.html

Method	Acc(%)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
PLM-Finetuned								
Bert	31.80%	-	-	-	-	-	-	-
RoBERTa	37.10%	-	-	-	-	-	-	-
Bart-base	-	23.00	-	-	10.35	44.33	24.29	20.80
Bart-large	-	26.37	-	-	11.65	44.92	24.34	21.75
PromptCLUE (T5)	-	18.75	-	-	6.65	40.88	21.90	18.31
LLM-Finetuned								
LLaMA	18.30%	29.25	-	-	16.46	45.88	26.57	23.31
ChatGLM	45.30%	31.10	-	-	18.94	43.94	31.48	29.39
Huatuo	28.60%	29.04	-	-	16.72	43.85	25.36	21.72
MedAlpaca	30.05%	16.35	-	-	9.78	44.31	27.05	24.55
LLM-Zero-shot								
Standard QA	45.83%	-	-	-	-	-	-	-
Chain-of-Thought	42.67%	34.02	24.27	18.03	13.57	46.20	20.39	21.97
ICP (Our method)	44.83%	42.73	30.43	22.57	16.89	45.24	20.84	23.40
LLM-Few-shot								
Standard QA	47.33%	-	-	-	-	-	-	-
Chain-of-Thought	50.00%	58.38	42.30	32.09	24.69	48.86	26.39	28.88
ICP (Our method)	**51.33%**	**59.84**	**43.68**	**33.38**	**25.86**	**49.89**	**27.43**	**29.70**

Table 2: Performance on Dataset CMExam.

4.2 Settings

The GPT 3.5-Turbo is selected as the primary model to evaluate, which includes 175B parameters and drives the online ChatGPT. All tests were conducted by calling OpenAI's official API. Unless specified, all experiments used the same parameters and were tested with the same version of the model (gpt-3.5-turbo-0613). To make the responses more deterministic and repeatable, we set the inference temperature to 0 to conduct greedy decoding, which always chooses the token with maximum probability instead of sampling. We keep the maximum context length (including prompt and response) of GPT 3.5-Turbo (4097 tokens) to avoid the potential performance penalty due to response length. The rest parameters are all set to default.

4.3 Evaluation Metrics

We evaluate model performance using both accuracy and NLG metrics. The accuracy generally represents the model's ability to conduct clinical reasoning. To assess the quality of the generated analysis, we employ BLUE [Papineni *et al.*, 2002] and ROUGE [Lin, 2004], which are commonly used in medical inference [Liu *et al.*, 2023; Tu *et al.*, 2023] or medical report generation [Liu *et al.*, 2021].

4.4 Baselines

To fully reveal the performance of LLMs with ICP, we evaluate several competitive baselines as follows.

- **PLM-Finetuned:** These baselines involve pre-trained language models (PLMs) extensively fine-tuned on the training set, including **BERT** [Vaswani *et al.*, 2017], **RoBERTa** [Liu *et al.*, 2019], **Bart-base** [Lewis *et al.*, 2019], **Bart-large**, and **PromptCLUE** [Zhang and Xu, 2022] (based on T5 model [Raffel *et al.*, 2020]). Encoder-only models like BERT and RoBERTa are designed to output results directly, whereas encoder-decoder models such as Bart and PromptCLUE can generate detailed explanations. Experimental setup and results referred to Liu *et al.*(2023).
- **LLM-Finetuned:** These baselines include LLMs fully fine-tuned on the training set, including both general LLMs (**LLaMA** [Touvron *et al.*, 2023] and **ChatGLM** [Du *et al.*, 2021]) and medical LLMs (**Huatuo** [Wang *et al.*, 2023a] and **MedAlpaca** [Han *et al.*, 2023]). They support the generation of inferences and final answers. Moreover, the medical LLMs have also undergone additional training on medical corpora and medical question datasets [Wang *et al.*, 2023a; Han *et al.*, 2023].
- **LLM-Zero-shot:** The **Standard QA** employs the standard question-and-answer prompting that instructs the LLM (GPT3.5) to respond directly to the predicted answer without any explanation, as "*Here is a multi-choice question about medical knowledge, please output the correct answer according to the question.*" The **Chain-of-Thought (CoT)** promoting was introduced by [Wei *et al.*, 2022] and [Kojima *et al.*, 2022], which guides the LLM to solve questions step by step and prompts it to generate a detailed analysis, as "*Here is a multi-choice question about medical knowledge, please analyze it in a step-by-step fashion and deduce the correct answer.*"
- **LLM-Few-shot:** We further enhance the reasoning capabilities of LLM through **Few-shot** approach, leveraging the in-context learning ability of LLMs. Given the context length constraints of LLMs, we randomly select six examples across a variety of clinical question types, including clinical case analysis, understanding of clinical knowledge, and medical computation tasks.

4.5 Result

As illustrated in Table 2 and Table 3, the proposed ICP outperforms baselines in both CMExam and CNMLE-Clinical, which significantly improves both the accuracy and quality of generated analysis. This verified the identified knowledge seeds encompass crucial information and thus efficiently guide the LLM to generate a convincing inference. Concurrently, we observe a strong coherence between accuracy and NLG metrics, which suggests that an improvement in the reasoning process aids in addressing clinical questions.

Effect of Generating Analysis. Compared to the QA-based method, the CoT-based method, although producing detailed explanations, often compromises the final accuracy. This is different from the findings in other domains

Method	Acc(%)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
LLM-Zero-shot								
Standard QA	51.00%	/	/	/	/	/	/	/
Chain-of-Thought	48.00%	12.47	8.75	6.32	4.57	37.31	15.60	16.76
ICP (Our method)	**53.33%**	**19.38**	**13.72**	**10.01**	**7.33**	**38.43**	**16.84**	**19.54**
LLM-Few-shot								
Standard QA	51.83%	/	/	/	/	/	/	/
Chain-of-Thought	54.83%	35.72	25.29	18.46	13.49	41.68	18.91	23.88
ICP (Our method)	**58.83%**	**35.78**	**25.47**	**18.69**	**13.74**	**42.21**	**19.47**	**24.28**

Table 3: Performance on Dataset CNMLE-Clinical.

Method	Acc(%)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Baichuan2-Chat (7B)								
CoT	42.83%	38.54	27.08	19.81	14.63	43.99	19.40	22.21
ICP	44.67%	52.75	36.92	26.94	19.81	44.47	20.95	24.88
CoT with Few-shot	49.83%	53.12	37.26	27.44	20.46	41.89	21.63	24.60
ICP with Few-shot	**50.00%**	**56.10**	**39.95**	**29.79**	**22.44**	**45.91**	**24.53**	**26.50**
GPT-4 (∼ 1.7T)								
CoT	72.00%	38.73	27.43	20.16	14.94	45.91	19.34	21.88
ICP	72.00%	43.21	31.13	23.24	17.50	47.25	21.99	24.56
CoT with Few-shot	76.00%	56.59	41.34	31.82	24.98	50.41	28.29	29.79
ICP with Few-shot	**78.00%**	**60.15**	**44.34**	**34.44**	**27.25**	**51.87**	**29.99**	**31.66**

Table 4: Performance of Different Base LLMs on CMExam.

where generating a detailed reasoning process prior to outputting an answer can substantially improve model performance, as seen in mathematical reasoning [Wei *et al.*, 2022; Kojima *et al.*, 2022]. This divergence in the medical domain might arise because of its intrinsic demand for precise medical knowledge. Without the guidance of reliable medical knowledge, CoT might inadvertently generate false knowledge or hallucinations, consequently impairing performance. Similarly, the previous evaluation of LLMs in medicine [Wu *et al.*, 2024b] also verified such a phenomenon.

Effect of Knowledge Seed. In contrast to CoT, our introduced ICP framework also improves the quality of generated explanations. Under the zero-shot setting, ICP achieves an increment of 3.32 in BLEU-4 and 1.43 in ROUGE-L for CMExam. Similarly, for CNMLE-Clinical, the uplift is 2.73 in BLEU-4 and 2.78 in ROUGE-L, resulting in accuracy improvements of 2.16 and 5.33, respectively. In CMExam with zero-shot, the accuracy of ICP is slightly lower than QA. This can be attributed to the intricacies of questions in CMExam, which involve more disciplines and usually contain more entities in longer analysis than CNMLE-Clinical (see Table 1). The incorporation of Few-shot strategy enables LLM to fully harness its in-context learning ability, which efficiently learns the patterns for problem-solving from examples, thus consistently improving model performance. The ICP with few-shot adeptly utilizes the predicted knowledge seeds for inference and achieves the best performance.

5 Analysis and Case Study

5.1 Impact of the LLM Size

To investigate the efficacy and generalizability of our framework, we conducted a comprehensive performance analysis on LLMs of varying sizes.

Baichuan2-chat (7B). The Baichuan2-chat-7B (BC2) was chosen to represent small LLMs, with its 7 billion parameters making it suitable for deployment on common PCs. It is an open-source bilingual LLM (Chinese and English) and has exhibited fine performance across various tasks. We assessed the reasoning abilities of BC2 under both zero-shot and few-shot scenarios for CoT and ICP using CMExam. Given BC2's extensive training on Chinese data, it achieved excellent performance on CMExam, only slightly behind GPT-3.5 (7B vs 175B). In a few-shot setting with ICP, its performance was nearly on par (50 vs 51.33), with a ROUGE-L score closely matching GPT-3.5 (26.50 vs 29.70). Such consistent and substantial improvements in a 7B model further validate the superiority of our framework, highlighting the potential of applying it to more resource-limited regions.

GPT-4 (∼1.7T). Subsequently, we examined GPT-4, which is unofficially estimated to possess around 1.7 trillion parameters, making it the most powerful model available in terms of both size and performance. We accessed the GPT-4 (version: 0613) via its official API with the same parameters as GPT-3.5. Due to the API access rate limitations, we randomly selected 50 samples for testing. The results underscored GPT-4's exceptional capabilities, as it achieved an accuracy of 72% even in zero-shot scenarios. Moreover, ICP consistently enhanced the model's accuracy and inference capabilities. This improvement also proves that our method can further enhance the model's ability for larger LLMs.

5.2 Performance on Different Medical Disciplines

To explore the usability of the LLMs integrated with ICP in medical scenarios, we further investigate its performance across different medical disciplines, areas of competency, and clinical departments. The subsequent analyses are based on the result of ICP under a few-shot setting with GPT3.5 in CMExam, as shown in Tables 5, 6, and 7.

As presented in Table 5, For clinical medicine, LLM with ICP achieved a high score of 60.77, qualifying for this licensing examination. This implies that its proficiency in medical knowledge is comparable to that of one certified physician. Furthermore, it achieved the highest ROUGE-L at 32.78 and BLEU-4 at 29.09. However, GPT-3.5 encounters challenges in public health and preventive medicine, yielding the lowest

Clinical Discipline	Count	Accuracy	ROUGE-L	BLEU-4
Clinical medicine	260	**60.77%**	**32.78**	**29.09**
Dentistry	95	43.16%	29.04	23.85
Pharmacy	74	54.05%	29.45	26.94
Traditional Chinese pharmacy	65	40.00%	23.44	21.12
Traditional Chinese medicine(TCM)	31	41.94%	26.06	20.74
Integrated TCM and Western medicine	49	44.90%	26.91	22.80
Public health and preventive medicine	26	30.77%	27.32	21.51

Table 5: The Performance on Different Clinical Disciplines.

Area of Competency	Count	Accuracy	ROUGE-L	BLEU-4
Disease diagnosis	221	53.85%	30.23	**26.86**
Medical knowledge	197	**54.31%**	30.17	25.58
Disease treatment	135	43.70%	27.86	24.91
Preventive medicine	40	52.50%	**30.57**	24.55

Table 6: Performance on Different Areas of Physician Competency.

Clinical Department	Count	Accuracy	ROUGE-L	BLEU-4
Internal medicine	118	56.78%	30.98	28.21
TCM	90	38.89%	23.68	19.47
General	67	50.75%	30.46	25.95
Dentistry	50	44.00%	30.98	26.81
Surgery	43	48.84%	32.92	28.95
Clinical laboratory	35	**71.43%**	28.91	25.36
Infectious diseases	31	51.61%	28.55	24.51
Obstetrics and gynecology	26	57.67%	29.77	27.15
Pediatrics	19	63.16%	31.17	26.67
Radiology	16	43.75%	**35.44**	**30.05**

Table 7: Performance on Different Clinical Departments.

accuracy of 30.77%. In contrast, BC2 achieved an accuracy of 50%. Given that GPT-3.5 is predominantly trained on English corpus, this disparity is primarily attributed to that it involves many Chinese healthcare policies and regulations, such as "*What is the stipulated standard for total coliform in China's drinking water sanitation guidelines?*".

To holistically evaluate a physician's capabilities, the examination integrates a variety of clinical questions, including an assessment of medical knowledge, disease diagnosis, disease treatment, and preventive medicine. As illustrated in Table 6, the LLM achieves an accuracy exceeding 52% in three areas. The disease treatment necessitates considerations of the main symptoms, disease history, current condition, and other factors, making it a more complex clinical reasoning. Therefore, more adaptations and improvements are required to cope with these questions.

Diving deeper into various clinical departments based on Table 7, it reveals that the model performance varies considerably across diseases, ranging from a low of 38.89% in TCM to a high of 71.43% in the clinical laboratory. This variation can be attributed to the inherent complexity differences among diseases; for instance, questions about the laboratory department primarily focus on the interpretation of biomarkers, offering a more concentrated context.

5.3 Error Analysis

To further investigate the role of the ICP framework in the reasoning process, we conducted an error analysis based on the results of ICP with Few-shot (GPT-3.5) in CMExam, as recorded in Table 8. We analyzed the metrics associated with the predicted knowledge seeds. Each ground-truth explanation contains nearly ten medical entities. When compared

	Count of KS	Precision of KS	Recall of KS	F1-score of KS	Length	ROUGE-L	BLEU-4
Correct	9.88	**0.312**	**0.285**	**0.266**	137.48	**31.09**	**27.21**
Incorrect	10.75	0.283	0.272	0.249	141.08	28.24	24.42

Table 8: Comparison of Generated Analysis by Answering Questions Correctly and Incorrectly. KS means knowledge seeds

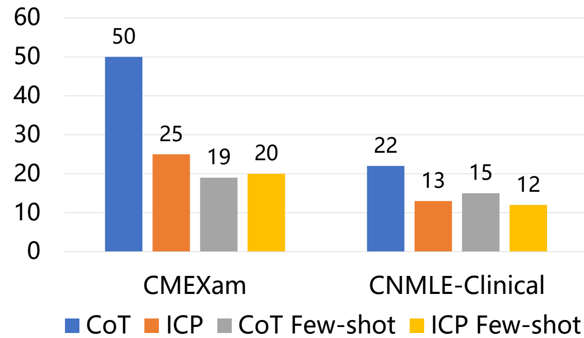

Figure 2: Questions Failed to Conclude An Answer.

with our predicted set of ten knowledge seeds, it became evident that questions with correct answers often had better knowledge seeds, as indicated by higher recall, precision, and F1 scores. Furthermore, regarding the length of responses and NLG metrics, we observed that the correct response typically has higher ROUGE and BLEU scores despite having a shorter average length compared to the incorrect one. Overall, this suggests that our designed ICP framework and mined knowledge seeds can guide LLMs to conduct focused discussions on potentially critical aspects for question-solving, thereby effectively enhancing the quality of generated analyses and improving their reasoning ability in the medical domain.

After reviewing the reasoning process, unlike the standard QA that directly provides an answer option, several analyses of CoT and ICP fail to conclude with an exact answer option, as highlighted in Figure 2. It suggests that there might be multiple correct answers or no correct option. While there are insightful discussions in the reasoning process, and some approximately reached the answer, the final answer is still uncertain due to the deficiency of medical knowledge or incomplete reasoning. This underscores the rigor and complexity of clinical reasoning, necessitating further exploration into integrating more knowledge or enhancing reasoning abilities.

6 Conclusion

In this study, we proposed a simple yet effective In-Context Padding framework that 1) identifies potential knowledge seeds using the medical knowledge graph; and 2) enhances LLM on clinical reasoning. Experiments have shown that our framework can significantly enhance the inference capabilities of LLMs. During the LLM reasoning process, the mined knowledge seeds effectively bridge the knowledge gaps between LLMs and the medical domain. Extensive ablation studies and error analysis have proven the robustness and generalizability of our framework. Our efforts aim to ensure that LLMs are effective and equitable in specialized domains, especially in healthcare, thereby promoting global health equity through artificial intelligence solutions.

Ethical Statement

While the explored clinical questions involve specific analysis of clinical cases, all cases have been anonymized and de-identified to ensure that no personal information is disclosed The primary objective of this research is to investigate the reasoning abilities of LLMs in the clinical domain. To that end, we opted for a comprehensive evaluation using clinical questions. However, the results and discussions of this research are purely for academic research and analysis and do not serve as actual medical suggestions. Therefore, there is no negative impact on human health.

Contribution Statement

Jiageng Wu and Xian Wu contributed equally to this paper. Correspondence to Jie Yang.

References

[Agrawal *et al.*, 2022] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.

[Ayers *et al.*, 2023] John W Ayers, Adam Poliak, and et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 2023.

[Bernstein *et al.*, 2023] Isaac A Bernstein, Youchen Victor Zhang, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open*, 6(8):e2330320–e2330320, 2023.

[Besta *et al.*, 2023] Maciej Besta, Nils Blach, et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

[Du *et al.*, 2021] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[Gao *et al.*, 2023] Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*, 2023.

[Gottret and Schieber, 2006] Pablo Enrique Gottret and George Schieber. *Health financing revisited: a practitioner's guide*. World Bank Publications, 2006.

[Haakenstad *et al.*, 2022] Annie Haakenstad, Caleb Mackay Salpeter Irvine, et al. Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 399(10341):2129–2154, 2022.

[Han *et al.*, 2023] Tianyu Han, Lisa C Adams, et al. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[Jeblick *et al.*, 2022] Katharina Jeblick, Balthasar Schachtner, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022.

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

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

[Lee, 2023] Hyunsu Lee. The rise of chatgpt: Exploring its potential in medical education. *Anatomical Sciences Education*, 2023.

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

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

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

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

[Liu *et al.*, 2021] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762, 2021.

[Liu *et al.*, 2023] Junling Liu, Peilin Zhou, et al. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*, 2023.

[Liu *et al.*, 2024] Xiaocong Liu, Jiageng Wu, and et al. Uncovering language disparity of chatgpt on retinal vascular disease classification: Cross-sectional study. *Journal of Medical Internet Research*, 26:e51926, 2024.

[Lyu *et al.*, 2023] Qing Lyu, Josh Tan, Mike E Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T Whitlow. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential. *arXiv preprint arXiv:2303.09038*, 2023.

[Montgomery Jr, 2018] Erwin B Montgomery Jr. *Medical reasoning: the nature and use of medical knowledge*. Oxford University Press, 2018.

[Nayak *et al.*, 2023] Ashwin Nayak, Matthew S Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P Weinfurt, and Kevin Schulman. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Internal Medicine*, 183(9):1026–1027, 2023.

[Nori *et al.*, 2023] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, and others J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[Singhal *et al.*, 2023] 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. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[Tarcar *et al.*, 2019] Amogh Kamat Tarcar, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai, and Dattaraj Rao. Healthcare ner models using language model pretraining. *arXiv preprint arXiv:1910.11241*, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, and et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Tu *et al.*, 2023] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

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

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[Wang *et al.*, 2023a] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

[Wang *et al.*, 2023b] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023.

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

[Wu *et al.*, 2024a] Jiageng Wu, Xiaocong Liu, and et al. Clinical text datasets for medical artificial intelligence and large language models—a systematic review. *NEJM AI*, page AIra2400012, 2024.

[Wu *et al.*, 2024b] Jiageng Wu, Xian Wu, and et al. Large language models leverage external knowledge to extend clinical insight beyond language boundaries. *Journal of the American Medical Informatics Association*, page ocae079, 2024.

[Wu *et al.*, 2024c] Jiageng Wu, Xian Wu, Yefeng Zheng, and Jie Yang. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *arXiv preprint arXiv:2403.06611*, 2024.

[Yang *et al.*, 2023] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[Yao *et al.*, 2022] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[Yao *et al.*, 2023] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[Young *et al.*, 2018] Meredith Young, Aliki Thomas, Stuart Lubarsky, Tiffany Ballard, David Gordon, Larry D Gruppen, Eric Holmboe, Temple Ratcliffe, Joseph Rencic, Lambert Schuwirth, et al. Drawing boundaries: the difficulty in defining clinical reasoning. *Academic Medicine*, 93(7):990–995, 2018.

[Zhang and Xu, 2022] Xuanwei Zhang and Liang Xu. Promptclue: A zero-shot learning model that supports full chinese tasks, September 2022.

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