MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs

Anonymous ACL submission

Abstract

Medical tasks such as diagnosis and treatment planning require precise and complex reasoning, particularly in life-critical domains. Unlike mathematical reasoning, medical reasoning demands meticulous, verifiable thought processes to ensure reliability and accuracy. However, there is a notable lack of datasets that provide transparent, step-by-step reasoning to validate and enhance the medical reasoning ability of AI models. To bridge this gap, we introduce **MedReason**, a large-scale high-quality medical reasoning dataset designed to enable faithful and explainable medical problem-solving in large language models (LLMs). We utilize a structured medical knowledge graph (KG) to convert clinical QA pairs into logical chains of reasoning, or "thinking paths", which trace connections from question elements to answers via relevant KG entities. Each path is validated for consistency with clinical logic and evidence-based medicine. Our pipeline generates detailed reasoning for various medical questions from 7 medical datasets, resulting in a dataset of **32,682** question-answer pairs, each with detailed, step-by-step explanations. Experiments demonstrate that fine-tuning with our dataset consistently boosts medical problem-solving capabilities, achieving significant gains of up to **7.7%** for DeepSeek-Ditill-8B. Our top-performing model, MedReason-8B, outperforms the Huatuo-o1-8B, a state-of-the-art medical reasoning model, by up to **4.2%** on the clinical benchmark MedBullets. We also engage medical professionals from diverse specialties to assess our dataset's quality, ensuring MedReason offers accurate and coherent medical reasoning. Our data, models, and code will be publicly available.

1 Introduction

Recent advancements in Reasoning Large Language Models (Xie et al., 2024b; Zhong et al., 2024; Wang et al., 2024a) highlight the remarkable effectiveness of utilizing *Chain-of-Thought (CoT)* reasoning (Huang and Chang, 2022; Miao et al., 2024) prior to the final answers. Although general-purpose reasoning models achieve parity with human performance in math and coding tasks (Guo et al., 2025; Team et al., 2025; Jaech et al., 2024), their applications in the medical domain have not been fully explored. One of the key challenges is the scarcity of high-quality CoT data, which is essential for developing medical reasoning models. Studies like s1 (Muennighoff et al., 2025) and

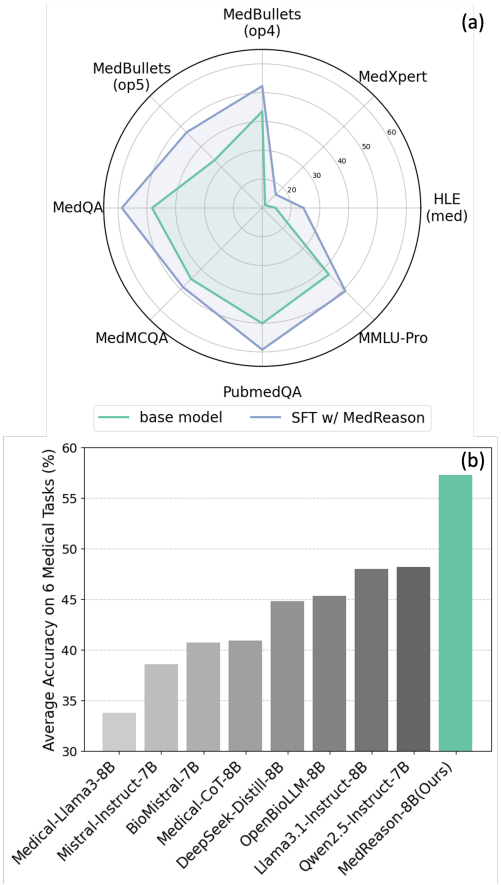

Figure 1: **MedReason-8B significantly enhances medical reasoning capability in LLMs.** (a) SFT with MedReason consistently improves base LLMs across multiple datasets. (b) Our fine-tuned model achieves state-of-the-art performance among 7-8B LLMs.

Question: Identify a treatment commonly associated with acute respiratory distress syndrome (ARDS) that has not been proven effective according to evidence-based research.
Ground-truth Answer: Inhaled nitric oxide

GPT-4o Generated Reasoning: ... taking all this into account, early steroid administration seems to be a common yet not strongly effective treatment for ARDS based on the evidence.

Ours Generated Reasoning: ... Inhaled nitric oxide can temporarily improve oxygenation in ARDS patients by addressing hypoxemia. However, evidence-based research indicates that it does not improve survival rates or long-term outcomes.

Figure 2: An Example from Huatuo's CoT Data, highlighting the factual error in the reasoning process generated by GPT-4o. In comparison, our generated reasoning leads to correct answer with accurate knowledge.

LIMO (Ye et al., 2025) have illustrated the crucial role of high-quality data in improving the LLMs' reasoning capability. The limited scalability of this high-quality medical data has hindered the development of more powerful medical intelligence.

One straightforward way to tackle this challenge is by distilling CoT data from open-source reasoning models (Madhusudhan et al., 2025; Huang et al., 2024; Min et al., 2024). However, these methods often omit **quality filtering** (i.e. verifying whether the generated CoT data logically leads to the correct answer). (Pang et al., 2025) introduces BOLT, which integrates LLMs within a multi-agent framework to produce extensive CoT data while employing an outcome reward model to filter out low-quality reasoning traces. Nevertheless, the proportion of medical-specific reasoning data remains limited, compromising the resulted model's clinical applicability. Efforts like HuatuoGPT-o1 (Chen et al., 2024b) aim to bridge this gap by generating medical CoT data using GPT-4o (Hurst et al., 2024). However, it is inevitable that general-purpose LLMs will generate responses that include factual errors. As shown in Fig. 2, GPT-4o erroneously concludes that early administering steroids is not a highly effective treatment for ARDS, contradicting established findings (Qadir et al., 2024).

In the medical domain, it is vital for the model to guarantee both the quality and rigorous factual guidance throughout every reasoning step, upholding medical reliability and clinical validity. In this paper, we argue that knowledge graph (KG) integration can provide **factual guidance** during CoT data generation, ensuring (1) logical coherence across all reasoning steps, and (2) clinical validity grounded in established medical knowledge. To achieve this, we propose a novel data generation pipeline that actively constrains the reasoning process to align with medical facts from KG, enhancing direct clinical utility. Specifically, our approach expands medical question-answering pairs into high-quality CoT data by searching reasoning paths from a high-quality medical KG (Chandak et al., 2023), which serves as reliable medical knowledge sources. We initially compile question-answering pairs from 7 medical datasets, encompassing general knowledge QA datasets as well as clinically challenging datasets which require complex reasoning. As shown in Fig. 3, for each question-answering pair, we firstly prompt an LLM to extract entities from the question and answer components, then map these entities to corresponding nodes in our medical knowledge graph through either exact matching or LLM-based similarity selection (Sec. 3.1.1). We subsequently identify all reasoning paths connecting the question and answer entities within the knowledge graph, and instruct the LLM to prune reasoning paths that do not pertain to the current question (Sec. 3.1.2). Finally, the remaining reasoning paths serve as structural scaffolds to guide the LLM in generating medically grounded CoT explanations, enhancing interpretability and reasoning quality (Sec. 3.2).

To ensure the quality of the generated CoT data, we implement a verification step where an LLM answers each question using the generated CoT. We systematically eliminate any CoT samples that fail to produce correct answers, ensuring only logically sound and clinically valid reasoning paths are retained (Sec. 3.2). As shown in Tab. 1, our proposed generation pipeline yields high-quality CoT samples with consistently improved reasoning quality across all evaluation metrics.

We assess the effectiveness of `MedReason` through supervised fine-tuning (SFT) on 1) instruction fine-tuned models (LLaMA 3.1-Instruct-8B (Grattafiori et al., 2024), Mistral-Instruct-7B (Jiang et al., 2023)) and 2) medical reasoning specialists (Medical-CoT-8B (Karataş, 2025), DeepSeek-Distill-8B (Guo et al., 2025)). Our extensive experimental evaluation on 7 QA benchmarks, encompassing 4 common medical benchmarks (MedQA (Jin et al., 2021), MedMCQA (Pal

Data Source	Quality Filtering	Medical Specific	Factual Guidance	Expert Checking
Distillation	✗	✗	✗	✗
BOLT	✓	✗	✗	✗
Huatuo-o1 CoT	✓	✓	✗	✗
MedReason(ours)	✓	✓	✓	✓

Table 1: **Comparison between Chain-of-Thought data sources.** MedReason provides high-quality medical CoT with factual guidance. Medical experts from seven departments assess the generated CoT data sampled from MedReason, further ensuring the quality of our data.

et al., 2022), MMLU-Pro (Wang et al., 2024b), and PubMedQA (Jin et al., 2019)), and 3 challenging clinical benchmarks (MedBullets (Chen et al., 2024a), MedXpert (Zuo et al., 2025), and Humanity's Last Exam (HLE) (Phan et al., 2025)) demonstrate the following key benefits of our generated CoT data: First, supervised fine-tuning (SFT) with our data yields consistent performance improvements across diverse base models and benchmarks. Notably, it enhances both instruction-tuned LLMs (Tab. 2) and specialized medical reasoning models (Tab. 3), with our best model achieving state-of-the-art performance among 7-8B parameter LLMs on challenging clinical benchmarks (Tab. 4 and Fig. 1). Second, MedReason produces higher-quality medical reasoning through knowledge-graph grounded generation, outperforming existing datasets (Chen et al., 2024b) in both automated metrics (Tab. 2) and expert evaluations conducted by physicians across seven clinical specialties (Fig. 4). Third, our approach enables superior clinical utility, as evidenced by side-by-side comparisons showing our model generates more factually precise and clinically supportive reasoning chains than competing approaches (Fig. 2 and Fig. 5).

2 Related Works

Reasoning with knowledge in LLMs. Recent reasoning large language models (LLMs) have demonstrated impressive performance in the math and coding domains (Guo et al., 2025; Team et al., 2025; Jaech et al., 2024; Li et al., 2025), prompting the need for analogous development in the medical domain (Goh et al., 2024; Lucas et al., 2024). However, training these models typically requires vast amounts of high-quality data that include intermediate reasoning steps (Shao et al., 2024; Guo et al., 2025). Since manually annotating such data is not scalable, they are often distilled from more powerful LLMs. This distillation process introduces unique challenges for tasks that demand factual knowledge, as LLMs can prone to generating hallucinations (Huang et al., 2023). This issue is even more pronounced in the medical domain, where even state-of-the-art LLMs struggle to provide high-quality and accurate reasoning (Griot et al., 2025; Chen et al., 2024c; Patel et al., 2005). To address these challenges, this work introduces MedReason, a medical reasoning dataset with high-quality CoT data designed to elicit factual-based and interpretable medical reasoning within LLMs.

LLM-Distilled Medical Datasets. Pre-training and fine-tuning medical LLMs demand extensive and high-quality datasets. Earlier research has primarily focused on gathering instruction-tuning data to imbue general domain LLMs with medical expertise (Xie et al., 2024a; Li et al., 2023). Recently, to improve the medical reasoning capability of LLMs, Chen et al. (2024b) introduced a medical CoT dataset by leveraging GPT-4o (Hurst et al., 2024) for strategy-based retrieval, yielding 20K question-answer pairs with complex CoT data. However, directly employing general domain LLMs to generate CoT data poses challenges in maintaining the integrity of medical knowledge at every reasoning step. Inspired by (Xie et al., 2024c), which obtains dependable information to assist in producing multigranular captions for medical images, our work employs a medical knowledge graph (Chandak et al., 2023) to provide factual guidance during the generation of medical CoT data.

3 Method

3.1 Retrieving Reasoning Paths from Knowledge Graph

In this section, we detail the process of retrieving reasoning paths from the knowledge graph (KG). In the data generation process, we define the Language Model as LLM and the knowledge graph as G. We utilize OpenAI GPT-4o (Hurst et al., 2024) as LLM and employ PrimeKG (Chandak et al.,

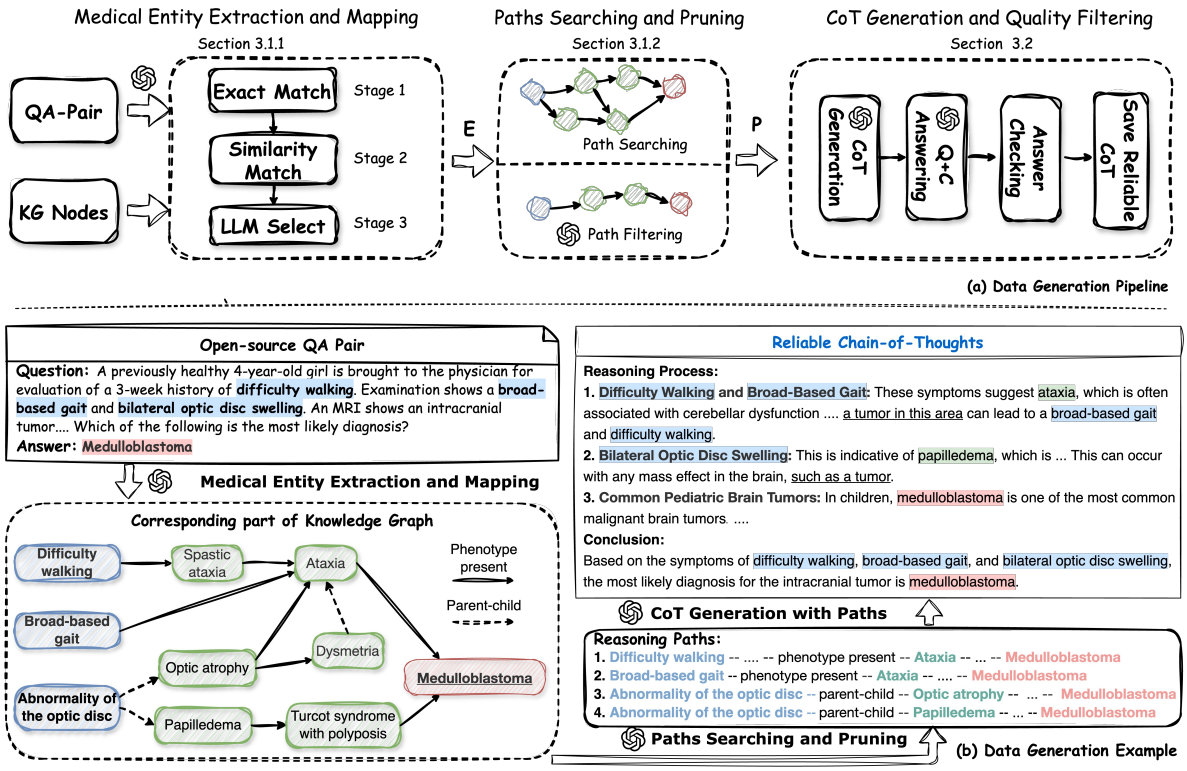

Figure 3: **Overview of Our Data Generation Pipeline.** We first extract and map entities within each medical Q&A pair, (see Sec. 3.1.1). Next, we search and prune the reasoning paths between Q&A entities in the KG (see Sec. 3.1.2), which are utilized as factual guidance to construct the CoT data(see Sec. 3.2). Finally, we discard any generated CoT that can not lead to reach the correct answer to ensure the quality of our data (see Sec. 3.2).

2023) as our knowledge base.

3.1.1 Medical Entity Extraction and Mapping

As shown in Fig. 3 (b), given a question Q and its corresponding answer A, we first utilize the Language Model LLM to identify the medical entities present in Q and A. This results in the extracted entity sets $\{e_i^Q\}_{i\in[n]}$ and $\{e_j^A\}_{j\in[m]}$, where n and m denote the number of entities in Q and A, respectively. These entities are then mapped to the corresponding nodes in the knowledge graph G through a three-step mapping process, as shown in Fig 3 (a). First, a text embedding model is used to encode each entity $e \in E$ and compute its similarity with the node embeddings in the G. This generates a ranked list of candidate matches, from which we then extract the Top-K most similar entities to form a candidate set S. Thirdly, we select the final entity from the Top-K entities, following the three matching stages:

Stage 1 (Exact Match): The algorithm iterates over S and checks for an exact match with e. If an exact match is found $e \in S$, the corresponding node is selected.

Stage 2 (Similarity Match): If an exact match is not found and the top similarity score exceeds a predefined threshold τ (set to 0.85 in our case), the most similar entity from S is selected.

$$\hat{e} = \arg\max_{s_k \in S} \cos(e, s_k), \text{ if } \cos(e, s_k) > \tau \quad (1)$$

Stage 3 (LLM-based Selection): If no suitable candidate is found in the above two stages, we instruct the LLM to analyze the question-answer context and the entity name to determine the most relevant node from S. The selection prompt I_{select} is demonstrated in Fig. 7 in the Appendix.

$$\hat{e} = LLM\left(S, Q, A \mid I_{\text{select}}\right), \quad (2)$$

Finally, we derive mapped entity sets from the graph, denoted as $\{\hat{e}_i^Q\}_{i\in[n]}$ and $\{\hat{e}_j^A\}_{j\in[m]}$, respectively. We detail the algorithm in Appendix Algorithm 1. As shown in Fig. 3(b), besides the entities *difficulty walking* and *broad-based gait* which exactly match graph nodes, the entity *bilateral optic disc swelling* is mapped to a similar concept, *Abnormality of the optic disc*, in the knowledge graph.

3.1.2 Paths Searching and Pruning

Given the mapped entities $\{\hat{e}_i^Q\}_{i\in[n]}$ from the question and $\{\hat{e}_j^A\}_{j\in[m]}$ from the answer, our goal is to

identify reasoning paths that logically connect the question to its corresponding answer. These paths will later serve as guidance for CoT generation (Sec. 3.2), ensuring that every reasoning step (1) originates from authoritative medical knowledge and (2) maintains factual consistency with the KG. Specifically, for each question–answer entity pair $\{\hat{e}_i^Q, \hat{e}_j^A\}$, we apply the graph-searching algorithm to extract reasoning paths that capture their most direct correlations. This set of reasoning paths for the node pair is represented as $\tilde{P}_{i,j}$. We demonstrate examples of paths in Fig. 3 (b), where the entities in question (blue nodes) are connected to the answer entity (red node). However, there may be a significant number of paths of the same length linking $\hat{e}_i^Q$ to $\hat{e}_j^A$ within the KG G. To ensure that we can retrieve the reasoning paths that are correlated to our question, we employ the LLM to prune the irrelevant paths as shown in Fig. 3 (a). In particular, we provide the shortest paths set $\tilde{P}_{i,j}$ and the question Q to the LLM, which is prompted to select K paths that are most correlated to the question. In summary, for the node pair $\{\hat{e}_i^Q, \hat{e}_j^A\}$, the reasoning paths searching and filtering process is denoted as:

$$\{\tilde{P}_{i,j}^k\}_{k\in[K]} = \texttt{shortest_paths}\left(\hat{e}_i^Q, \hat{e}_j^A, G\right),$$
$$P_{i,j}^k = LLM\left(\tilde{P}_{i,j}^k, Q \mid I_{\text{prune}}\right), \quad k \in [K] \tag{3}$$

where I_{prune} is the path pruning prompt. More details can be found in Appendix Fig. 8. Finally, we aggregate all identified reasoning paths across the question-answer pairs, forming the complete set $\mathcal{P} = \{P_{i,j}^k\}_{i\in[n],j\in[m],k\in[K]}$, for CoT generation. As illustrated in Fig. 3 (b), we explore various reasoning paths that connect question and answer entities within the corresponding subgraph of the KG. After filtering out the irrelevant paths, the pruned paths effectively identify the link between symptoms like 'Difficulty walking' and the diagnosis 'Medulloblastoma', uncovering the critical intermediary disease 'Ataxia'.

3.2 CoT Generation and Quality Filtering

CoT Generation with Reasoning Paths Utilizing step-by-step reasoning paths in $\mathcal{P}$ as guidance, we are able to distill the reliable knowledge from the off-the-shelf KG into our CoT data. To achieve this, we prompt the LLM to analyze the given reasoning paths and elaborate on the relevant ones to formulate medically grounded CoT explanations of the response, represented as:

$$C = LLM\left(Q, A, \mathcal{P} \mid I_{\text{gen}}\right), \tag{4}$$

where I_{gen} represents our carefully designed generation prompt (see Appendix Fig. 9). As illustrated in Fig. 3b, our approach produces clinically grounded reasoning chains - for instance, beginning with symptom analysis, progressing through pathological deduction (e.g., tumor identification), and ultimately concluding the final diagnosis. Each step maintains direct alignment with the KG-derived evidence in $\mathcal{P}$, ensuring both factual accuracy and clinical relevance.

Quality Filtering Lastly, to ensure the quality of the generated data, we design a simple quality filtering strategy to filter out low-quality CoT data. Specifically, for each generated CoT C, we prompt the LLM to produce an answer $\hat{A}$ using *only* the information contained in C (see details in Appendix Fig. 10). The answer generation process can be denoted as:

$$\hat{A} = LLM\left(Q, C \mid I_{\text{eval}}\right), \tag{5}$$

where I_{eval} denotes the prompt for generation. Subsequently, as illustrated in Fig. 3 (a), we compare $\hat{A}$ with the original ground-truth answer A. We apply this quality filtering strategy to all 45K generated samples, retaining only those CoT instances (32K) that yield correct answers, to ensure both logical validity and factual accuracy in our final dataset.

4 Experiment

4.1 Experimental Setup

Data Collection and Preprocessing. Our data pipeline generates CoT reasoning for Question-Answering (QA) pairs. To curate medical QA pairs, we gather datasets such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubmedQA (Jin et al., 2019), MMLU (Hendrycks et al., 2020), HuatuoGPT-o1 (Chen et al., 2024b), MedXpert (Zuo et al., 2025), and Humanity's Last Exam (HLE) (Phan et al., 2025). To prevent data leakage, we exclusively use the training set from each dataset for CoT data generation, culminating in a total of 55K QA pairs. We exclude the QA pairs that cannot produce CoT data using our pipeline due to the absence of entities in either the question or answer, such as when the answer is merely a number, ultimately leading to 45K QA pairs for CoT data generation. Additional statistics for the

Benchmarks	Llama3.1-Instruct-8B			Mistral-Instruct-7B		
	original	w/ huatuo CoT	w/ ours	original	w/ huatuo CoT	w/ ours
MedQA	58.7	**70.2 (+11.5)**	68.4 (+9.7)	48.2	**59.9 (+11.7)**	58.7 (+10.5)
MedMCQA	56.0	**58.2 (+2.2)**	57.5 (+1.5)	44.9	46.9 (+2.0)	**48.9 (+4.0)**
PubmedQA	75.2	76.1 (+0.9)	**77.6 (+2.4)**	50.1	57.5 (+7.4)	**59.2 (+9.1)**
MMLU-Pro	58.2	59.9 (+1.7)	**63.1 (+4.9)**	42.7	47.6 (+4.9)	**50.8 (+8.1)**
MedBullets(op4)	48.7	53.3 (+4.6)	**57.5 (+8.8)**	43.5	50.0 (+6.5)	**52.3 (+8.8)**
MedBullets(op5)	42.5	49.7 (+7.2)	**52.3 (+9.8)**	33.4	46.1 (+12.7)	**47.1 (+13.7)**
MedXpert	13.2	**17.3 (+4.1)**	16.4 (+3.2)	11.4	14.4 (+3.0)	**16.6 (+5.2)**
HLE (med)	13.6	14.6 (+1.0)	**16.5 (+2.9)**	14.6	14.6 (+0.0)	**24.3 (+9.7)**
Avg	45.8	49.9 (+4.1)	**51.2 (+5.4)**	36.1	42.1 (+6.0)	**44.7 (+8.6)**

Table 2: **Results of instruction-tuned LLMs fine-tuned with Huatuo complex CoT and `MedReason` (Ours).** Integrating our high-quality reasoning data into supervised fine-tuning significantly improves model performance across various benchmarks and model types.

generated and quality-filtered datasets are provided in Table 7 in the Appendix.

Baseline Models. To evaluate the effectiveness of our dataset across various base models, we select multiple 7-8B models as baselines and conduct supervised fine-tuning (SFT) on `MedReason`. Specifically, we fine-tune two representative instruction-tuned models: LLaMA 3.1-Instruct-8B (Grattafiori et al., 2024) and Mistral-Instruct-7B (Jiang et al., 2023). Following (Chen et al., 2024b), we train these models for three epochs using a learning rate of 5e-6 and a batch size of 128, employing DeepSpeed-ZeRO stage 3 (Rajbhandari et al., 2020). To further assess the impact of our dataset on reasoning models, we also fine-tune Medical-CoT-8B (Karataş, 2025), DeepSeek-Distill-8B (Guo et al., 2025), and Huatuo-o1-RL-8B (Chen et al., 2024b) using the same hyperparameter settings.

Finally, in Tab. 4, we benchmark our model against three categories of models: (1) General LLMs, including LLaMA 3.1-Instruct-8B (Grattafiori et al., 2024), Mistral-Instruct-7B (Jiang et al., 2023), and Qwen-Instruct-7B (Yang et al., 2024); (2) Medical-Specific LLMs, such as Medical-Llama (Qiu et al., 2024), OpenBioLLM (Ankit Pal, 2024), Huatuo-o1-SFT (Huang et al., 2024), and BioMistral (Labrak et al., 2024); and (3) Medical Reasoning Models, including Medical-CoT (Karataş, 2025) and Huatuo-o1-RL (Huang et al., 2024).

Benchmarks. We evaluate on standard medical benchmarks: MedQA (USMLE test set) (Jin et al., 2021) , MedMCQA (validation set) (Pal et al., 2022), MedBullets (Chen et al., 2024a), health and biology tracks of MMLU-Pro (Wang et al., 2024b), and PubMedQA (test set) (Jin et al., 2019). Additionally, we evaluated the medical sections of some challenging LLM benchmarks, including the expert-level medical knowledge and advanced reasoning benchmark MedXpert (Zuo et al., 2025), and the most recent medical questions in Humanity's Last Exam (HLE) (Phan et al., 2025). We report the pass@1 accuracy for all benchmarks.

4.2 Results

`MedReason` on Instruction Fine-tuned Model. In this section, we showcase the enhancement of Instruction Finetuned Models using `MedReason`. Tab. 2 displays the accuracy (%) of Llama3.1-Instruct-8B and Mistral-Instruct-7B across various medical benchmarks. The model finetuned on `MedReason` (**w/ ours**) consistently outperforms both the base models and the Huatuo CoT data (Chen et al., 2024b) finetuned model.

For Llama3.1-Instruct-8B, finetuning using `MedReason` improves the average accuracy from **45.8%** to **51.2%** (**+5.4%**), surpassing the **+4.1%** gain achieved with Huatuo CoT data. In the case of Mistral-Instruct-7B, the improvement is even more substantial, rising from **36.1%** to **44.7%** (**+8.6%**), exceeding Huatuo CoT data's **+6.0%** gain. Our method demonstrates consistent and substantial improvements, particularly in complex reasoning tasks (HLE) and specialized clinical benchmarks (MedBullets, MedXpert).

`MedReason` on Reasoning Models. We then showcase the improvements in reasoning models achieved through fine-tuning with our `MedReason` dataset. As shown in Tab. 3, fine-tuning with our dataset (w/ ours) significantly enhances

Base Model	Data	Clinical Challenging Datasets				
		MedBullets(op4)	MedBullets(op5)	MedXpert	HLE (med)	Avg
Medical-CoT-8B	original	39.3	34.1	12.6	15.5	25.4
	w/ ours	**49.0** (+9.7)	**41.9** (+7.8)	**14.2** (+1.6)	**17.5** (+2.0)	**30.6** (+5.3)
DeepSeek-Distill-8B	original	41.9	35.1	13.5	11.7	25.5
	w/ ours	**53.6** (+11.7)	**49.0** (+14.0)	**15.9** (+2.4)	**14.6** (+2.9)	**33.3** (+7.7)

Base Model	Data	Common MedicalQA Datasets				
		MedQA	MedMCQA	PubmedQA	MMLU-Pro	Avg
Medical-CoT-8B	original	49.0	42.6	68.0	48.7	52.1
	w/ ours	**58.0** (+9.0)	**46.6** (+4.0)	**74.6** (+6.6)	**50.4** (+1.7)	**57.4** (+5.3)
DeepSeek-Distill-8B	original	55.4	49.0	**73.9**	53.8	58.0
	w/ ours	**63.7** (+8.3)	**51.8** (+2.8)	73.0 (-0.9)	**57.5** (+3.8)	**61.5** (+3.5)

Table 3: **Fine-tuning with our data further enhances reasoning LLMs.** We perform supervised fine-tuning on reasoning models in both general and medical domains, with our data consistently improving performance by providing high-quality medical knowledge.

Model	MedBullets (op4)	MedBullets (op5)	MedXpert	MedQA	MedMCQA	PubmedQA	Avg
Llama3.1-Instruct-8B	43.2	40.9	14.3	58.7	56.0	75.2	48.0
Qwen2.5-Instruct-7B	50.0	41.6	12.6	57.0	55.6	72.7	48.2
Mistral-Instruct-7B	43.5	33.4	11.4	48.2	44.9	50.1	38.6
Medical-Llama3-8B	33.4	25.3	9.0	40.3	46.8	48.0	33.8
OpenBioLLM-8B	39.2	35.7	10.7	57.7	54.1	74.1	45.3
BioMistral-7B	46.4	33.1	12.4	45.0	40.2	66.9	40.7
Medical-CoT-8B	39.3	34.1	12.6	49.0	42.6	68.0	40.9
DeepSeek-Distill-8B	41.9	35.1	13.5	55.4	49.0	73.9	44.8
Huatuo-o1-SFT-8B	53.3	49.7	17.3	70.2	58.2	76.1	54.1
Huatuo-o1-RL-8B	55.2	51.3	16.7	**72.6**	60.4	79.2	55.9
MedReason-8B (**ours**)	**57.5**	**55.5**	**19.0**	71.8	**60.7**	**79.4**	**57.3**

Table 4: **Comparison across various medical benchmarks.** Our best model achieves state-of-the-art performance comparing with 7B-8B LLMs.

performance across both clinical and general medical question-answering tasks compared to the original model without fine-tuning. On challenging clinical datasets, Medical-CoT-8B achieves an average gain of 5.3%, while DeepSeek-Distill-8B demonstrates an even greater improvement of 7.7%. Similarly, for general medical QA datasets, Medical-CoT-8B improves by 5.3% on average, and DeepSeek-Distill-8B gains 3.5%. These results highlight the effectiveness of our KG-driven dataset MedReason in enhancing reasoning LLMs by integrating factual-guided medical knowledge.

MedReason achieves state-of-the-art on 7B-8B LLMs. Finally, we show that applying MedReason fine-tuning further enhances the advanced RL-ed reasoning LLM, Huatuo-o1-RL-8B. By fine-tuning Huatuo-o1-RL-8B with MedReason, we obtain a new state-of-the-art model, termed MedReason-8B. The results in Tab. 4 show that MedReason-8B outperforms all other models across five evaluation datasets, achieving the highest average score of **57.3%**. It surpasses its base model, Huatuo-o1-RL-8B, by 1.4%, demonstrating the effectiveness of MedReason fine-tuning. Notably, it demonstrates even greater improvements on clinical reasoning tasks, outperforming Huatuo-o1-RL-8B by 4.2% on MedBullets (op5) and 2.3% on MedXpert. Compared to other strong baselines, such as OpenBioLLM-8B and DeepSeek-Distill-8B, MedReason-8B demonstrates substantial performance gains of approximately 12%. These results confirm that our KG-guided CoT data enhances reasoning capabilities, establishing a new state-of-the-art for medical QA tasks.

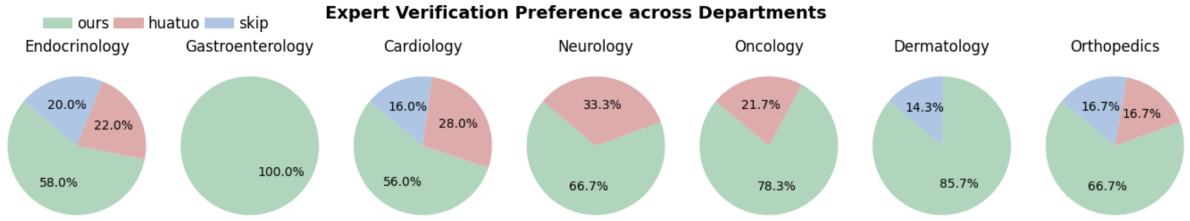

Figure 4: **Expert verification on CoT data quality** across seven medical specialties. Doctors were shown CoT data from both our dataset (green) and Huatuo-Complex CoT (red) and asked to pick the higher-quality explanation or skip (blue) if neither was clearly superior. Each pie indicates the proportion of doctors in that specialty who selected our data, Huatuo-Complex CoT, or skipped. Our CoT data was consistently favored across all the specialties.

Benchmark	w/o QF	w/ QF
MedQA	0.669	**0.684**
MedMCQA	0.569	**0.575**
PubmedQA	0.737	**0.776**
MedBullets (op4)	0.571	**0.575**
MedBullets (op5)	0.510	**0.523**
MedXpert	**0.178**	0.164
Avg	0.539	**0.550**

Table 5: **The ablation study on quality filtering (QF).** Transposed version: datasets as rows, methods as columns.

4.3 Ablation Study and Expert Verification

Effect of Quality Filtering. As outlined in Section 3.2, we implement quality filtering to ensure the generated CoT data effectively guides a LLM to produce correct answers. To evaluate its impact, we perform an ablation study using the LLama3.1-Instruct-8B model. Table 5 illustrates that quality filtering enhances performance across the majority of medical datasets, increasing the average score by 1.1%. This further underscores the crucial role of CoT data quality in enhancing the reasoning capabilities of LLMs in medical applications.

CoT Case study. To showcase the effectiveness of our dataset, we analyze a challenging medical case from MedBullets, comparing its performance against two other reasoning models—DeepSeek-distilled and HuatuoGPT-o1—in Fig. 5 (see Appendix). Our dataset accurately diagnosed Cri-du-Chat syndrome, linking the patient's symptoms to a chromosome 5p deletion, consistent with the ground-truth answer, thus validating its reasoning precision. In contrast, DeepSeek-distilled model frequently responded with *I'm not sure*, as indicated by the yellow shadow in Fig. 5, reflecting uncertainty and a lack of clarity, which makes it unsuitable for assisting doctors. While HuatuoGPT-o1 provided a more confident response, it relied on incorrect knowledge, as indicated by the red shadow, leading to an inaccurate diagnosis.

Expert Verification. To further evaluate the medical accuracy and clarity of CoT data within MedReason, we engage domain experts from seven different medical specialties. Each expert is given two anonymized CoT data, one produced by MedReason and the other by HuatuoGPT-o1, and asked to choose which explanation, if any, is more accurate and easier to understand. Experts could also skip if they found both CoT data equally good or inadequate. For each specialty, we evaluated responses to over 25 randomly sampled questions. For detailed procedure, see Appendix Sec. A.7. Fig. 4 presents the results, illustrating the proportion of high-quality CoT data sources verified by licensed clinicians. As shown, specialists across all departments prefer MedReason, with Gastroenterology experts unanimously selecting MedReason (100%), and Dermatology and Oncology also show a strong preference (over 80%). These findings suggest that MedReason consistently provides more medically precise and coherent reasoning across multiple specialties, underscoring its potential for clinical applications.

5 Conclusion

This paper presents a scalable, knowledge-guided pipeline for automatically generating high-quality medical Chain-of-Thought (CoT) data. By leveraging structured knowledge graphs to anchor the reasoning process, our method produces medically grounded and interpretable explanations that enhance the clinical validity of LLM-generated reasoning. Experiments on instruction-tuned and reasoning-specialized LLMs demonstrate consistent improvements across medical benchmarks, particularly in complex clinical scenarios, while expert evaluations confirm superior reasoning quality over prior methods. We hope our work can encourage future exploration on clinically valid reasoning to advance trustworthy medical AI.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024c. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1 others. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Emre Karataş. 2025. Deepseek-r1-medical-cot. https://huggingface.co/emredeveloper/DeepSeek-R1-Medical-COT.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Sathwik Tejaswi Madhusudhan, Shruthan Radhakrishna, Jash Mehta, and Toby Liang. 2025. Millions scale dataset distilled from r1-32b. https://huggingface.co/datasets/ServiceNow-AI/R1-Distill-SFT.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1):148.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, and 1 others. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Bo Pang, Hanze Dong, Jiacheng Xu, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2025. Bolt: Bootstrap long chain-of-thought in language models without distillation. *arXiv preprint arXiv:2502.03860*.

Vimla L Patel, José F Arocha, and Jiajie Zhang. 2005. Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning*, 14:727–750.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Nida Qadir, Sarina Sahetya, Laveena Munshi, Charlotte Summers, Darryl Abrams, Jeremy Beitler, Giacomo Bellani, Roy G Brower, Lisa Burry, Jen-Ting Chen, and 1 others. 2024. An update on management of adult patients with acute respiratory distress syndrome: an official american thoracic society clinical practice guideline. *American journal of respiratory and critical care medicine*, 209(1):24–36.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Jun Wang, Meng Fang, Ziyu Wan, Muning Wen, Jiachen Zhu, Anjie Liu, Ziqin Gong, Yan Song, Lei Chen, Lionel M Ni, and 1 others. 2024a. Openr: An open source framework for advanced reasoning with large language models. *arXiv preprint arXiv:2410.09671*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024a. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. 2024b. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and 1 others. 2024c. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, and 1 others. 2024. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Appendix

A.1 Limitations

Our study focuses on using Knowledge Graphs (KGs) to guide chain-of-thought generation in text-based medical question answering. We do not consider extensions to multi-modal tasks, such as visual question answering or medical image analysis, which we leave for future work, but believe such extensions can be incorporated naturally. Furthermore, Reinforcement Learning with Verifiable Rewards (RLVR) is an effective post-training approach for language models. We will explore how to leverage KGs to provide dense rewards during RL, which we expect to be a promising direction for enhancing LLM performance on more clinically challenging tasks.

A.2 Use Of AI Assistant

We employed large language models (LLMs) in the dataset construction pipeline to extract medical entities, generate Chain-of-Thought from reasoning paths, and conduct quality checking.

During the preparation of this manuscript, we used OpenAI's GPT-4.1 model for minor language refinement and smoothing of the writing. The AI tool was not used for generating original content, conducting data analysis, or formulating core scientific ideas. All conceptual development, experimentation, and interpretation were conducted independently without reliance on AI tools.

A.3 Data Statistics

The following table demonstrates the specification of the data distribution of `MedReason` across various source datasets. In the case of Humanity's Last Exam (HLE) (Phan et al., 2025) and MedXpert (Zuo et al., 2025), we allocated 57 and 666 data samples, respectively, for training and reserved the rest for testing. For other datasets, we strictly use only the training data for data generation to avoid any potential data leakage.

A.4 Prompts

The prompts used in our data generation pipeline are shown in the following figures. We provide details about prompts for (a) identifying entities in the question and answer (Fig. 6), (b) selecting the most relevant nodes for extracted entities (Fig. 7), (c) pruning irrelevant paths (Fig. 8), (d) CoT data generation with paths (Fig. 9), and (e) generating the answer based on provided CoT data for quality filtering (Fig. 10).

A.5 Detailed Algorithm

In algorithm 1, we present our detailed algorithm for entity extraction and mapping, as outlined in Sec. 3.1.1. First, a text embedding model is used to encode each entity $e \in E$ and compute its similarity with the node embeddings in the G. This generates a ranked list of candidate matches, from which we then extract the Top-K most similar entities to form a candidate set S. Thirdly, we select the final entity from the Top-K entities, following the three matching stages:

Stage 1 (Exact Match): The algorithm iterates over S and checks for an exact match with e. If an exact match is found $e \in S$, the corresponding node is selected.

Stage 2 (Similarity Match): If an exact match is not found and the top similarity score exceeds a predefined threshold τ (set to 0.85 in our case), the most similar entity from S is selected.

$$\hat{e} = \arg\max_{s_k \in S} \cos(e, s_k), \text{ if } \cos(e, s_k) > \tau \quad (6)$$

Stage 3 (LLM-based Selection): If no suitable candidate is found in the above two stages, we instruct the LLM to analyze the question-answer context and the entity name to determine the most relevant node from S. The selection prompt I_{select} is demonstrated in Fig. 7 in the Appendix.

$$\hat{e} = LLM\left(S, Q, A \mid I_{\text{select}}\right), \quad (7)$$

Finally, we derive mapped entity sets from the graph, denoted as $\{\hat{e}_i^Q\}_{i\in[n]}$ and $\{\hat{e}_j^A\}_{j\in[m]}$, respectively. As illustrated in Fig. 3, besides the entities *difficulty walking* and *broad-based gait* , which exactly match graph nodes, the entity *bilateral optic disk swelling* is mapped to a similar concept, *Abnormality of the optic disc*, in the knowledge graph. In this study, we utilize data sources from publicly published researches, which have already passed the ethic review.

A.6 Detailed comparison between `MedReason`-8B and Huatuo-o1-RL-8B.

In Fig. 11, we present a comprehensive comparison of `MedReason`-8B and Huatuo-o1-RL-8B in

Clinical Scenario: A newborn boy is evaluated in the hospital nursery 2 hours after birth... The patient's mother received routine prenatal care and the pregnancy was uncomplicated. The patient's birth weight is 2.6 kg (5 lb 11 oz), which is at the 5th percentile. His height and head circumference are in the 15th and 3rd percentile, respectively. His temperature is 97.1°F (36.2°C), blood pressure is 57/42 mmHg, pulse is 140/min, and respirations are 38/min. On physical exam, the patient has a wide nasal bridge, down slanting palpebral fissures, and widely spaced eyes. He has good respiratory effort with a high-pitched cry. This patient is most likely to have experienced a deletion on which of the following chromosomes?

Options: (A) 4p; (B) 5p; (C) 5q; (D) 7q; (E) 15q.

Ground-truth Answer: 5p

Sample from [Medbullets]

DeepSeek-distilled Qwen

Uncertain Reasoning

Okay, so I have this medical case to solve… He has good respiratory effort with a high-pitched cry. The cry can sometimes be a clue, but I'm not sure. High-pitched cry is sometimes associated with conditions like croup or maybe something else, but I'm not sure how that ties into the genetic aspect. … Or maybe 7q, which can cause a wide bridge of the nose. Alternatively, 5p deletions can cause a phenotype with a triangular face, short stature, and maybe some heart issues. But I'm not sure about the facial features. … In summary, the facial features point towards 7q deletion.

HuatuoGPT-o1

Inaccurate Knowledge

Alright, let's think this through. We've got a newborn boy here, and he's showing some distinct physical features. He's got a wide nasal bridge, down-slanting eyes, and widely spaced eyes. ... Oh right, Williams syndrome! That's the one. People with Williams syndrome often have these facial features, and it's usually linked to a deletion on chromosome 7. So, that's a big clue ... And chromosome 5p is associated with Cri-du-chat syndrome, which has a unique cat-like cry, but this baby's cry is described as high-pitched, not like a cat. So that's not it either. ...

I'm pretty confident that the answer is chromosome 7q.

Our Model

Accurate Reasoning

Finding Reasoning Paths: 1. … 2… 3. Dysmorphic features are characteristic of certain chromosomal deletions, such as 5p deletion syndrome (cri-du-chat syndrome). 4. High-pitched cry is a hallmark of cri-du-chat syndrome, which is associated with a deletion on chromosome 5p. **Reasoning Process:** ... Exclusion of Other Syndromes: - Prader-Willi syndrome and Angelman syndrome are associated with deletions on chromosome 15q11-q13, but they do not present with a high-pitched cry, or the specific dysmorphic features seen in this patient. **Conclusion:** The most likely diagnosis for this patient is Cri-du-chat syndrome, which is associated with a deletion on chromosome 5p.

Figure 5: **Case Study on Medbullets Benchmark.** Some part of reasoning is omitted due to the space limitation. Our model generates accurate reasoning with reliable knowledge.

tackling complex clinical issues. Huatuo-o1-RL-8B delivered an incorrect answer due to overlooking the link between vomiting and hyperglycemia, whereas our MedReason-8B accurately identified the correct answer, grounded in factual reasoning.

A.7 Detail of Expert Instruction

In this section, we describe the expert verification procedure, in which multiple licensed doctors independently evaluated the quality of the generated reasoning. As shown in Tab. 6, each expert was presented with a pair of anonymized reasoning traces—one from Huatuo and one from MedReason, in a randomized order, and was asked to select the reasoning with higher overall quality. The evaluation criteria focused on both reasoning correctness and factual accuracy; experts could also choose to skip if the two reasonings were judged to be equivalent. For instances where experts provided conflicting preferences, we conducted an additional consensus stage to reconcile differing opinions and determine a final label, or assign a skip if consensus could not be reached. This multi-stage evaluation protocol helps ensure reliable, high-quality expert judgments.

Step	Expert Verification Procedure
1. Pairing & Blinding	For each medical question, we prepare two reasoning traces: one from HUATUO and one from MedReason(ours). The two reasonings are anonymized and presented to doctors in a *randomized order* to prevent source bias.
2. Preference Voting	Doctors are asked to select the higher-quality reasoning based on correctness and factualness. If the two reasonings are judged as equally good (or equally inadequate), the doctor selects **Skip**.
3. Disagreement Resolution	For question instances where the doctors' votes *disagree*, we conduct a follow-up round by returning these instances to the experts and requesting a final **agreement** decision (i.e., a resolved preference or skip).

Table 6: **Expert verification protocol for comparing reasoning quality between HUATUO and MedReason (ours).**

Datasets	MedQA	MedMCQA	PubmedQA	MMLU	MedXpert	Huatuo	HLE	Total
Raw	9595	9131	24826	1089	1000	9271	159	55071
Generated	8528	7598	20613	893	951	7010	132	45725
Quality Filtered	8016	6197	10444	827	666	6475	57	32682

Table 7: **Statistics of the collected QA dataset (Raw), Generated dataset, and final Quality-Filtered dataset.**

Prompt for identifying entities in question and answer

```
messages = [ {"role": "system", "content": "You are an AI assistant that helps people find information."},
{"role": "user", "content": """ You are a helpful, pattern-following medical assistant.
  Given the text in a medical or biomedical context, precisely extract all entities from the text.

  ### Output Format
  Strictly follow the JSON structure below.
  The type of each entity MUST STRICTLY BELONG to one type from:
  1. gene/protein 2. drug  3. effect/phenotype  4. disease  5. biological_process  6. molecular_function  7.
  cellular_component  8. exposure  9. pathway  10. anatomy

  ```json
  {{
  "Entity": [
    {{"id": "1", "type": "some_type", "name": "entity_name"}},
    {{"id": "2", "type": "some_type", "name": "entity_name"}},
  ]
  }}
  ```
"""}
]
```

Figure 6: Prompt for identifying entities in the question and answer.

Prompt for finding most related node entity

```
messages = [ {"role": "system", "content": "You are an AI assistant that helps people find information."},
{"role": "user", "content": """ You are a helpful, pattern-following medical assistant.
  Given a medical question and corresponding answer, an query entity which is extracted from the
  question, and a list of similar entities.
  Select ONE most correlated entity from the list of similar entities based on the question and query
  entity.
  SELECTED ENTITY MUST BE IN THE SIMILAR ENTITIES.
  IF there is not suitable entity in the similar entities, directly return the NONE.

  ### Output Format
  Strictly follow the JSON structure below:
  ```json
  {{
    "selected_entity": {{
      "name": "selected_entity_name",
      "id": a int number, the index of the selected entity in the similar entities list, from 0 to N-1
      "reason": "reason for choosing this entity"
    }}
  }}
  ```
"""}
]
```

Figure 7: Prompt for selecting the most relevant nodes for extracted entities.

Prompt for pruning irrelevant paths

```
messages = [ {"role": "system", "content": "You are an AI assistant that helps people find information."},
{"role": "user", "content": """ You are a helpful, pattern-following medical assistant.
  Given a medical question and possible relation paths that link to the answer.
  Select up to {topK} most correlated entity from the relation paths list based on the question and the
  answer.
  If total number of paths is less than {topK}, select all of them.

  ### Output Format
  Strictly follow the JSON structure below.
  ```json
  {{
  "Paths": [
    {{"ranking": "1", "path": "sample_path_1", reason": "reason for choosing this path"}},
    .....
  ]
  }}
  ```
"""}
]
```

Figure 8: Prompt for pruning irrelevant paths.

Prompt for Chain-of-Thoughts generation with paths

```
messages = [ {"role": "system", "content": "You are an AI assistant that helps people find information."},
{"role": "user", "content": """ You are an expert in the medical domain.
  Given a medical question, a set of reasoning paths, and a provided answer, your task is to reason step
  by step as if you are independently determining the possible reasoning paths and deriving the correct
  answer without prior knowledge of the given answer.

  1. Explore the question and pretend to generate multiple plausible reasoning paths. If any of the
  provided paths seem useful, incorporate them naturally as if you discovered them yourself.
  2. If none of the given paths seem relevant or correct, ignore them and generate your own reasoning
  approach based on your expertise.
  3. Analyze and evaluate the reasoning paths carefully, expanding on the most relevant ones to
  construct a logical, well-supported explanation.
  4. Do not mention the existence of predefined reasoning paths or the provided answer in your response.
  5. Do not assume the given answer is correct. Instead, determine the answer solely based on your
  reasoning.
  6. If your final conclusion contradicts the given answer, acknowledge potential discrepancies (e.g., "Wait,
  there might be something wrong") and refine your response accordingly.

  ### Input:
  Question: {question}
  Answer: {answer}
  Paths: {paths}
  ### Output:
  Finding reasoning paths:
  (you "discover" potential reasoning paths yourself by using the given paths if useful or generating your
  own if not. It should be concise as a list of paths)
  Reasoning Process:
   (Step-by-step reasoning process, do not assume the given answer is correct and do not mention the
  existence of answer.)
  Conclusion:
  (The final answer derived from your reasoning.)
"""}
]
```

Figure 9: Prompt for Chain-of-Thoughts generation with paths.

Prompt for generating answer with Chain-of-Thoughts reasoning

```
messages = [ {"role": "system", "content": "You are an AI assistant that helps people find information."},
{"role": "user", "content": """ You are an expert in the medical domain. You need to answer the following
  question based on the provided reasoning.
  YOU MUST USE THE PROVIDED REASONING TO ANSWER THE QUESTION.
  If the answer choices are provided, please choose ONE answer from the answer choices, ensuring your
  response concludes with the correct option in the format: 'The answer is A.'.\n{question}\n{option_str}
  Question:
  {question}
  {options}
  Reasoning:
  {reasoning}
"""}
]
```

Figure 10: Prompt for generating the answer based on provided CoT data.

Algorithm 1 Mapping Extracted Entities to Knowledge Graph Nodes

Require: Extracted entity set $\mathcal{E}$ from Q or A, Knowledge Graph G, text embedding model f_T, similarity threshold τ, lower case operation `lower`, prompt instruction $\mathcal{I}_s$ and LLM function f_L.

Ensure: Mapped knowledge graph nodes for each entity in $\mathcal{E}$

```
for each entity e in E do
    S ← get_similar_entities(e, G, f_T)
    s_top ← top_similarity(S)
    selected ← None
    for each candidate c in S do                          ▷ Stage 1: Exact Match
        if lower(e) = lower(c) then
            selected ← c
            break
        end if
    end for
    if selected = None and s_top > τ then                 ▷ Stage 2: Similarity Match
        selected ← S[0]
    end if
    if selected = None then                               ▷ Stage 3: LLM-based Selection
        selected ← f_L(Q, A, e, S|I_s)
    end if
    Map e to node selected ê
end for
```

Clinical Scenario: A 32-year-old female with obesity, hypertension, and type 2 diabetes seeks treatment for ongoing hyperglycemia. Her glycated hemoglobin is 7.9% (normal: 4.3–5.6%), and she shows consistent elevated post-meal blood glucose levels without any hypoglycemic episodes. Her current medications include metformin 1000 mg twice daily and lisinopril 5 mg daily, a regimen she has maintained for several years.\n\nShe has developed diabetes-related retinopathy and neuropathy, but her kidney function remains intact. Her symptoms include early fullness after eating, post-meal nausea, and frequent vomiting of undigested food consumed hours earlier. To manage these symptoms, she has adopted a pattern of eating small, frequent meals. She also experiences tingling sensations in both legs.\n\nWhich of the following medication modifications would be most appropriate to improve her blood glucose control?

Options: (A) Add acarbose; (B) Add dapagliflozin; (C) Switch to dulaglutide; (D) Add semaglutide; (E) Add pioglitazone; (F) Replace metformin with glipizide; (G) Add sitagliptin.

Ground-truth Answer: **Add dapagliflozin**

Sample from [MedXpert]

HuatuoGPT-o1 Inaccurate Knowledge

Thinking\n\nAlright, let's think about this. We've got a 32-year-old woman dealing with obesity, hypertension, and type 2 diabetes. She's been on metformin and lisinopril for a while, but her blood sugar levels are still high, especially after meals. She's also experiencing some unpleasant symptoms like feeling full quickly, nausea, and vomiting after eating, which makes sense because those are classic signs of gastroparesis. Oh, and she's got tingling in her legs, which could be related to her diabetes.\n\nNow, what do we do about her blood sugar? We need something that works well with her current meds and helps with her symptoms. Metformin is doing its job, but it's not enough on its own. We need to add something that targets post-meal spikes and helps with her symptoms. Hmm, what about acarbose? It's known for slowing down the digestion of carbs, which should help with the nausea and vomiting by giving her stomach more time to process food. Plus, it's designed to tackle those pesky post-meal glucose spikes.\n\nLet's think about the other options. Dapagliflozin and empagliflozin are SGLT2 inhibitors, which are great for lowering blood sugar but don't really address the nausea and vomiting. Dulaglutide, semaglutide, and liraglutide are GLP-1 receptor agonists, which can help with blood sugar and weight, but they might make nausea worse. Pioglitazone is more for improving insulin sensitivity, not really for nausea. And then there's glimepiride and sitagliptin, which are sulfonylureas and DPP-4 inhibitors, respectively. They work on increasing insulin but don't help with gastroparesis symptoms.\n\nSo, acarbose seems like the best fit here. It's designed for situations just like hers, where you've got high blood sugar and symptoms of gastroparesis. It's a good match because it can help manage her symptoms and improve her blood sugar control without making her feel worse. Yeah, adding acarbose makes the most sense.\n\n## Final Response\n\n**The answer is A.**

Our Model Accurate Reasoning

Finding Reasoning Paths: ### Finding Reasoning Paths:\n1. **Hyperglycemia and Type 2 Diabetes**: The patient has chronic hyperglycemia and type 2 diabetes, managed with metformin and lisinopril. Additional glucose control measures are needed.\n2. **Retinopathy and Neuropathy**: These are complications of diabetes, which are not directly related to blood glucose control but may necessitate a more aggressive glucose management strategy.\n3. **Frequent Symptoms of Hyperglycemia**: Post-meal nausea, vomiting, and early fullness suggest poor glucose control, possibly due to inadequate insulin or medications.\n4. **Lipid Profile**: A high LDL cholesterol and low HDL cholesterol indicate metabolic derangement, which may require additional medications to improve lipid control.\n5. **Tingling in Legs**: This could be due to neuropathy or other complications, but it is not directly linked to blood glucose management.\n6. **Medication Options**: The patient is already on metformin and lisinopril. Additional medications targeting glucose metabolism or insulin secretion may be considered.\n\n---\n\n### Reasoning Process:\n1. **Current Glucose Control**: The patient's glycated hemoglobin is 7.9%, which is significantly elevated. Metformin and lisinopril are her current medications. Metformin is effective for hyperglycemia but may not be sufficient in cases of advanced diabetes or when combined with other complications like retinopathy and neuropathy.\n\n2. **Symptoms of Hyperglycemia**: The symptoms of post-meal nausea, vomiting, and early fullness are classic for poorly controlled hyperglycemia. These symptoms suggest that her blood glucose levels are not being tightly controlled, possibly due to inadequate insulin or medications.\n\n3. **Treatments for Hyperglycemia**: To improve blood glucose control, medications targeting insulin secretion, glucose uptake, or insulin resistance may be considered. Examples include sulfonylureas, DPP-4 inhibitors, GLP-1 receptor agonists, and GLP-1 receptor antagonists.\n\n4. **Medication Options**:\n - **DPP-4 Inhibitors (e.g., dapagliflozin)**: These medications improve insulin secretion and glucose uptake, which can help control hyperglycemia. They are often used in diabetic patients with retinopathy and neuropathy.\n - **GLP-1 Receptor Antagonists (e.g., semaglutide)**: These medications inhibit GLP-1 secretion, promoting insulin secretion and glucose uptake. They are particularly effective in patients with advanced diabetes and retinopathy.\n - **GLP-1 Receptor Agonists (e.g., sitagliptin)**: These medications stimulate GLP-1 secretion, which can help lower blood glucose levels but are less commonly used in advanced diabetes.\n - **Sulfonylureas (e.g., glimepiride, glipizide)**: These medications stimulate insulin secretion and are often used in diabetic patients. However, they are generally considered less effective in advanced diabetes compared to DPP-4 inhibitors or GLP-1 receptor antagonists.\n\n5. **Combination Therapy**: In patients with advanced diabetes, combining DPP-4 inhibitors and GLP-1 receptor antagonists (e.g., dapagliflozin and semaglutide) can provide synergistic effects to improve blood glucose control.\n\n---\n\n### Conclusion:\nThe most appropriate medication modification to improve her blood glucose control is **to add dapagliflozin**. This treatment targets insulin secretion and glucose uptake, addressing the symptoms of poorly controlled hyperglycemia and supporting her overall glucose management.\n\n## Final Answer\n\n**(B) Add dapagliflozin**

Figure 11: Full comparison between MedReason-8B and Huatuo-o1-RL-8B.