MedCoach: Enhancing Medical Reasoning in LLMs via Knowledge Graph-Augmented Chain-of-Thought Distillation

Chuan Li[1], Ye Lyu[1], Chengyu Wang[2], Mingyuan Fan[1], Cen Chen[1]*
[1]East China Normal University, Shanghai, China
[2]Alibaba Group, Hangzhou, China
{lichuan, yelyu, mingyuan_fmy}@stu.ecnu.edu.cn
chengyu.wcy@alibaba-inc.com, cenchen@dase.ecnu.edu.cn

Abstract

Despite the advanced capabilities of Large Language Models (LLMs), training specialized reasoning models for the medical domain remains a significant challenge due to the scarcity of high-quality, large-scale Chain-of-Thought (CoT) data. Moreover, the intermediate reasoning steps in teacher-generated CoT data can be redundant and noisy, leading models to acquire spurious patterns and resulting in suboptimal performance. To address these issues, we propose MedCoach, a novel framework that introduces a dedicated coach role to guide the student model through question decomposition, thereby smoothing its learning curve in medical reasoning. The framework employs a curriculum-oriented warm-up on simplified sub-questions, facilitating domain adaptation before advancing to complex long-chain reasoning. To ensure the fidelity of the intermediate chain-of-thought signals, we augment this phase with medical knowledge graphs to suppress factual drift and mitigate reasoning noise at a granular level. Subsequently, we introduce a targeted factual perturbation mechanism to foster fine-grained discrimination between valid fact utilization and subtle factual misapplications. Extensive experiments across diverse benchmarks demonstrate notable improvements over existing methods, validating the effectiveness of MedCoach. Code is available at https://github.com/DIaacKr/MedCoach.

1 Introduction

Reasoning constitutes a cornerstone of medical expertise, underpinning critical processes across the healthcare landscape, from diagnostic decision-making and treatment planning to biomedical research and patient communication (Kassirer, 2010). Recent advances in large language models (LLMs), exemplified by DeepSeek-R1 (Guo et al., 2025) and GPT-o1 (Jaech et al., 2024), have demonstrated substantial gains in reasoning capabilities, particularly in domains such as mathematics and programming (Wei et al., 2022; Guan et al., 2025; xio; Wang et al., 2025b; Guo et al., 2024). However, extending these advances to medical applications remains non-trivial (Chen et al., 2025b; hongzhou yu et al., 2025; Huang et al., 2025b; Xu et al., 2025). A key obstacle lies in the stringent demand for high-quality chain-of-thought (CoT) data in medicine, where verifying the factual soundness of intermediate reasoning steps is inherently difficult.

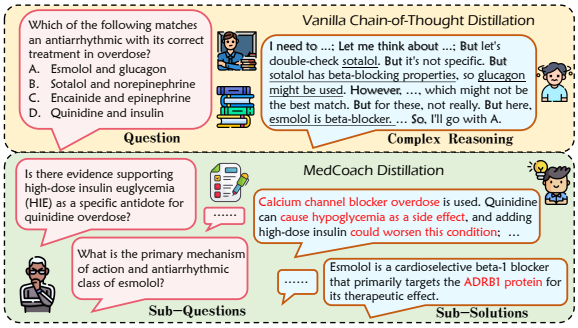

Figure 1: Comparison between vanilla KD and our method. Shaded parts represent the structural complexity in medical CoTs, and underlined parts indicate knowledge-related content. Red annotations denote the corresponding grounded refinements.

Moreover, knowledge distillation (KD) has emerged as a promising paradigm for transferring reasoning capabilities from high-capacity teacher models to smaller student models (Hinton et al., 2015; Xu et al., 2024; Zhou et al., 2023). However, as illustrated in Figure 1, the vanilla KD method, which requires students to directly replicate teacher-generated reasoning trajectories, often fails in practice due to a fundamental mismatch in cognitive capacity. Generally, small models lack the representational strength to internalize complex, multi-step reasoning processes from the outset. Furthermore, due to the presence of redundant or erroneous inter-

* Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 33724–33743

mediate steps in CoT data, students may inadvertently preserve or even exaggerate these artifacts, leading to spurious or hallucinatory outputs. Such issues not only degrade the fidelity of distilled reasoning but also compromise trustworthiness, especially in safety-critical domains such as medicine, where factual reliability are indispensable.

In this paper, we propose MedCoach, a coach-guided distillation framework that bridges the reasoning gap between teacher and student models in medicine. Unlike standard CoT distillation, MedCoach introduces a mediating coach that deconstructs complex reasoning from the teacher into a sequence of sub-questions, each of which is grounded in factual knowledge retrieved from a medical knowledge graph. This decomposition can produce a fine-grained, high-quality learning trajectory, enabling students to acquire reasoning skills incrementally through targeted supervision on knowledge-grounded sub-solutions. In addition to decomposition, we enrich reasoning trajectories with benign paraphrasing and adversarial revision of intermediate steps, yielding paired valid and erroneous variants that explicitly contrast desired versus undesired reasoning paths.

Through comprehensive experiments across diverse configurations, including varying teacher and student model architectures and benchmarks, our approach demonstrates significant and notable improvements over existing test-time scaling techniques and state-of-the-art baselines. The principal contributions of this work are as follows:

- To the best of our knowledge, this work is the first study to focus on multi-step reasoning in medical LLMs via distillation techniques integrated with knowledge graphs.
- We propose a systematic, fine-grained CoT distillation framework. Operating at the granularity of self-derived sub-questions and augmented by a knowledge graph and multi-strategy perturbations, this framework is designed to enhance the factual reliability and reasoning robustness of student models in complex medical domains.
- We conduct extensive large-scale experiments across diverse medical reasoning benchmarks. Our results demonstrate superior effectiveness and generalization compared to existing distillation approaches.

2 Related Work

Medical LLMs. With continuous technological advancements, LLMs have demonstrated remarkable capabilities not only in general domains but have also increasingly made significant inroads into the medical field (Zhang et al., 2023; Li et al., 2023; Bao et al., 2023; Zhang et al., 2024; Yang et al., 2024c; Wei et al., 2025; Ye et al., 2023; Yang et al., 2024b; Yano et al., 2025; Wang et al., 2025a). Following the advent of reasoning models like OpenAI's GPT-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), there has been a growing emphasis within medical applications on pursuing enhanced reasoning capabilities (Chen et al., 2025b; hongzhou yu et al., 2025; Jiang et al., 2025; Huang et al., 2025b; Xu et al., 2025). While specific reasoning-focused models have emerged, achieving strong medical reasoning often comes at a substantial computational cost, frequently overlooking the critical scenario of resource-constrained settings.

Knowledge Distillation for Reasoning. The rise of LLMs has spurred considerable research activity in knowledge distillation, which is recognized as a key technique for efficiently improving model performance under relatively low-resource conditions (Hinton et al., 2015; Xu et al., 2024; Gou et al., 2021). Given the proprietary nature of many state-of-the-art models, black-box distillation (i.e., distilling knowledge using only model inputs and outputs) remains the predominant approach in practice (Jiao et al., 2019; Wang and Yoon, 2021; Hsieh et al., 2023; Yue et al., 2024). The emergence of reasoning-specialized LLMs has consequently driven increased attention toward CoT-based distillation techniques (Magister et al., 2023; Li et al., 2022; Wang et al., 2023; Chen et al., 2024; Cai et al., 2025). Conventional CoT distillation methods often overlook factuality enhancement and tolerate the generation of hallucinations.

Knowledge Graphs for LLMs. Knowledge Graphs (KGs) are valued for their ability to ensure strong factual accuracy and traceability, making them a highly promising mechanism for augmenting LLMs (Wu et al., 2024; Jia et al., 2025; Mondal et al., 2024; Zhao et al., 2025; Sun et al., 2024). MedReason (Wu et al., 2025b) leverages KGs by retrieving relevant paths pertinent to a query to construct CoT demonstrations, thereby strengthening medical reasoning capabilities. Wang et al. (2024a) introduced Chain-of-Knowledge (CoK) prompting, a method that improves reasoning fac-

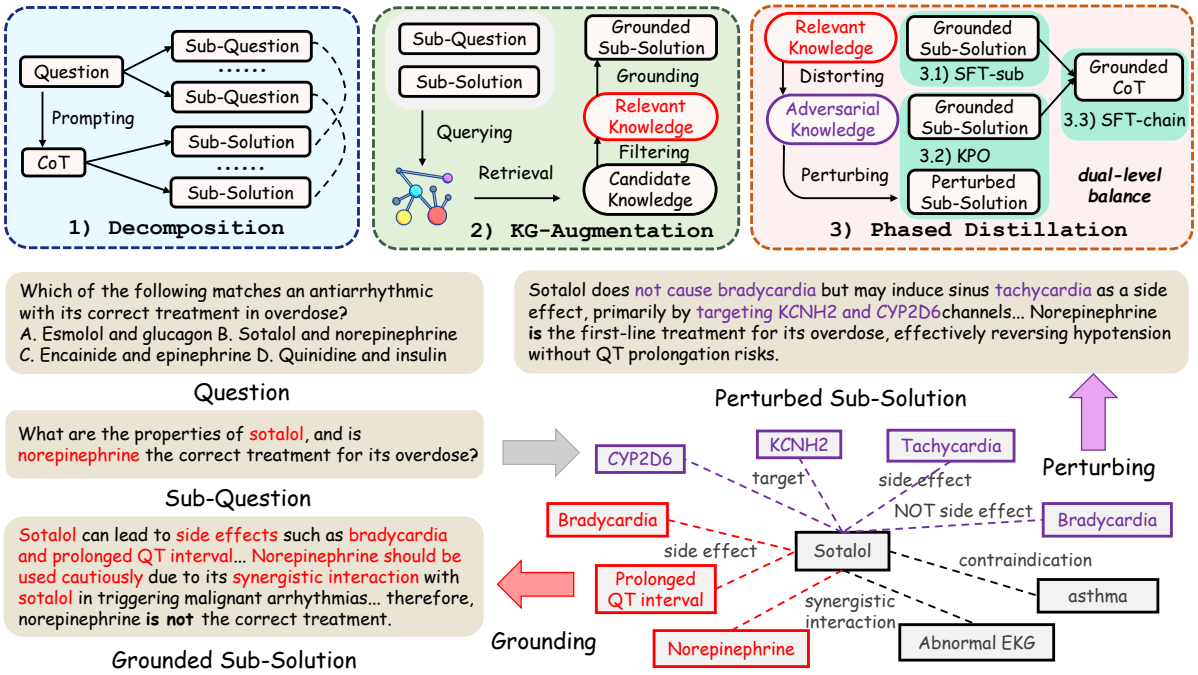

Figure 2: A brief overview of the **MedCoach** framework. Red highlights indicate relevant and beneficial knowledge retrieved from the KG, whereas purple highlights denote adversarial and misused knowledge generated to enhance model robustness during distillation.

tuality by guiding language models to generate structured, verifiable evidence triples as the basis for their rationale. To address the challenges of annotation dependence and poor generalization in semantic parsing, Huang et al. (2025a) introduced TARGA, which automatically synthesizes targeted question-query pairs on the fly for a given query for use as effective in-context learning examples. Wu et al. (2025a) present a framework for augmenting corpus-based RAG by using a knowledge graph to simulate spreading activation for more effective multi-source retrieval and fact-enhanced answer generation. However, these prior efforts lack extension to general-purpose QA paradigms, or simply leverage KGs for unidirectional factual augmentation without incorporating fine-grained discrimination between correct and erroneous or misapplied knowledge.

3 Methodology

We propose a concise distillation framework (see Figure 2) designed to enhance the reasoning capabilities of lightweight models. The framework consists of three components: (i) decomposition splits complex tasks into fine-grained sub-questions to create granular training scenarios; (ii) knowledge graph augmentation enriches the decomposed sub-questions to improve factual reliability; and (iii) phased CoT distillation, which operates in three sequential stages. In the first stage, sub-question-level supervised fine-tuning uses the decomposed sub-questions and teacher demonstrations to provide preliminary training on moderately difficult samples. The second stage, sub-question-level preference optimization, leverages both grounded and perturbed samples to improve the model's discriminative ability in knowledge utilization. Finally, chain-level supervised fine-tuning employs knowledge-enhanced reasoning chains to promote coherent global reasoning.

3.1 Decomposition

Question Decomposition Our framework first acquires standard CoT distillation data. Notably, with the availability of powerful open-source reasoning models, such CoT data can be obtained by direct prompting:

$$\mathcal{C}_i = \mathcal{F}_{\text{CoT}}(\mathcal{B}_i, Q_i) \quad (1)$$

where $\mathcal{F}_{\text{CoT}}$ denotes the teacher reasoning process that generates the full reasoning trace $\mathcal{C}_i$ given the medical context $\mathcal{B}_i$ and question Q_i. We then introduce reasoning-aware decomposition to structurally unpack complex medical questions using this initial CoT. The hierarchical distillation process begins by analyzing the reasoning trace $\mathcal{C}_i$ to

generate medically meaningful sub-questions:

$$\mathcal{S}_i = \mathcal{F}_{\text{decom}}\left(\mathcal{B}_i, Q_i, \mathcal{C}_i\right) \tag{2}$$

where $\mathcal{F}_{\text{decom}}$ denotes the decomposition function implemented via an LLM (i.e., the coach), which takes the original question as input and generates a set of sub-questions. This is achieved by using a structured prompt (see Appendix I for this and all subsequent prompt templates) that explicitly instructs the model to decompose the task and return the sub-questions in a parseable format. A detailed description of the pipeline can be found in Appendix H. The output is then post-processed to extract the sub-question set: $\mathcal{S}_i = \{q_{i,j}\}_{j=1}^{M}$, which contains M medically coherent sub-questions that systematically unpack the complexity of Q_i. Crucially, this decomposition targets reasoning-critical questions where medical knowledge integration or interpretive ambiguity may occur, such as drug side effects and synergistic relationships (see Figure 2).

Solution Segment Alignment To enable fine-grained knowledge augmentation, we establish an exact correspondence between sub-questions $q_{i,j}$ and their supporting reasoning segments in $\mathcal{C}_i$ through an alignment function using the coach:

$$c_{i,j} = \mathcal{F}_{\text{align}}(q_{i,j}, \mathcal{C}_i) \tag{3}$$

where the coach is prompted to identify text spans in $\mathcal{C}_i$ that directly address each $q_{i,j}$. The output is then validated by exact string matching against the original reasoning path $\mathcal{C}_i$, retaining only those sub-solutions with verifiable matches. This containment check ensures that $c_{i,j}$ consists exclusively of contiguous text from $\mathcal{C}_i$ that semantically aligns with $q_{i,j}$, enabling subsequent medical knowledge operations at a precisely matched granularity.

3.2 Knowledge Graph Augmentation

Efficient Knowledge Retrieval Conventional knowledge enhancement paradigms typically depend on intricate entity linking and subgraph generation pipelines (Wu et al., 2025b; Li et al., 2025b; Wu et al., 2025a), which introduce significant computational overhead. We circumvents these limitations through direct semantic retrieval that leverages contextual coherence at the sub-question level. The retrieval mechanism operates as:

$$\mathbf{e}_c = f_{\text{emb}}(\text{concat}(q_{i,j}, c_{i,j})) \tag{4}$$

$$\mathcal{T}_{i,j} = \underset{t \in \mathcal{G}}{\text{top-}K}\ \left(\phi(\mathbf{e}_c, f_{\text{emb}}(t))\right) \tag{5}$$

where f_{emb} denotes a medical text embedding model (we use MedEmbed-large-v0.1 (Balachandran, 2024), optimized for medical semantics), ϕ represents the similarity metric (cosine similarity by default), and concat$(\cdot)$ performs contextual fusion. Prior to embedding computation, each knowledge triple $t = (h, r, o)$ drawn from the PrimeKG knowledge graph (Chandak et al., 2023) is deterministically textualized via domain-specific templates $t \mapsto \Psi(h, r, o)$. For instance, the template *"{drug} is clinically indicated for the treatment of {disease}."* yields concrete sentences such as *"Aspirin is clinically indicated for the treatment of cardiovascular disease."* Leveraging the limited combinatorial space of medical entity-relation pairs in $\mathcal{G}$, this lightweight textualization achieves a favorable balance between computational efficiency and semantic fidelity, as detailed in Appendix F.

This approach achieves substantial efficiency gains while maintaining retrieval precision. Furthermore, operating at the sub-question granularity focuses retrieval on clinically actionable knowledge elements, such as drug-disease contraindications and biomarker-diagnosis associations, rather than broad conceptual relations.

Knowledge-Grounded Reasoning Refinement To bolster factual grounding while preserving medical coherence, retrieved knowledge undergoes relevance filtering before integration. Leveraging the teacher model's domain expertise, we perform selective knowledge retention:

$$\mathcal{T}_{i,j}^{\text{rel}} = \mathcal{F}_{\text{filter}}(\mathcal{T}_{i,j}, q_{i,j}, c_{i,j}) \tag{6}$$

where $\mathcal{F}_{\text{filter}}$ implements relevance assessment, retaining only triples relevant to the specific context. This ensures subsequent augmentation operates on medically relevant knowledge. Thus, the refinement is conducted through conditional rewriting:

$$\tilde{c}_{i,j} = \mathcal{F}_{\text{ground}}(c_{i,j}, \mathcal{T}_{i,j}^{\text{rel}}, \mathcal{B}_i) \tag{7}$$

where $\mathcal{F}_{\text{ground}}$ denotes the integration function implemented via the coach, which conditionally revises the reasoning segment using the retrieved knowledge. The grounding process maintains factual consistency with medical knowledge, contextual alignment with background information $\mathcal{B}_i$, and preservation of the original solution logic. By anchoring medical statements to verifiable knowledge sources, this approach substantially reduces confabulation risks while respecting existing valid

reasoning. Therefore, the original CoT $\mathcal{C}_i$ is reorganized into a knowledge-enhanced CoT by substituting in the original CoT, yielding $\tilde{\mathcal{C}}_i = \{\tilde{c}_{i,1}, \ldots, \tilde{c}_{i,M}\}$ through segment-wise refinement.

3.3 Phased Chain-of-Thought Distillation

Sub-Question-Based Supervised Fine-Tuning Here, the student model learns to solve atomic medical sub-questions with integrated knowledge. Training pairs consist of contextualized sub-questions and their augmented sub-solutions: $\mathcal{D}_{\text{sub}} = \{(\mathcal{B}_i, q_{i,j}, \tilde{c}_{i,j})\}_{i,j}$. The optimization objective maximizes the likelihood of generating the correct sub-solutions:

$$\mathcal{L}_{SFT_{sub}} = -\sum_{i=1}^{N}\sum_{j=1}^{M_i} \log P\left(\tilde{c}_{i,j} \mid \mathcal{B}_i, q_{i,j}; \theta\right) \quad (8)$$

where θ denotes the student parameters and M_i is the sub-question count for case i. This phase establishes precise knowledge grounding capabilities.

Knowledge-Aware Preference Optimization To cultivate precise medical knowledge discrimination capabilities, we propose the Knowledge-Aware Preference Optimization (KPO) technique with strategically designed adversarial samples. This paradigm addresses a critical challenge in medical reasoning: semantically similar knowledge representations may lead to medically divergent outcomes. Inspired by Li et al. (2025a), for each knowledge-augmented sub-solution $(\tilde{c}_{i,j})$, we construct negative samples through triple distortion that preserves superficial similarity while introducing medically significant deviations:

- *Type I (Thematic distractions):* Replace original triples with semantically related but actually irrelevant knowledge. This simulates reasoning misdirection, where surface-level thematic associations overshadow actual medical applicability (e.g., focusing on protein binding sites when contraindications are needed for prescribing).
- *Type II (Entity corruptions):* Maintain related relationships but substitute critical entities with medically similar alternatives. This mimics entity misidentification errors, where practitioners correctly identify the required relationship types but confuse semantically similar entities (e.g., focusing on tranexamic acid when aminocaproic acid overdose requires urgent intervention).
- *Type III (Relationship inversions):* Preserve core entities while reversing critical medical relationships. This represents an insidious knowledge misuse scenario, where fundamental biomedical principles are misremembered (e.g., confusing contraindications with indications for high-risk medications).

The adversarial triples $\mathcal{T}_{i,j}^{\text{adv}}$ are processed through a similar rewriting mechanism in Eq. 7 to generate negative responses $c_{i,j}^{\text{neg}}$:

$$c_{i,j}^{\text{neg}} = \mathcal{F}_{\text{perturb}}(\mathcal{B}_i, q_{i,j}, \tilde{c}_{i,j}, \mathcal{T}_{i,j}^{\text{adv}}) \quad (9)$$

yielding semantically similar yet medically erroneous solutions that mirror authentic medical reasoning failures. Specifically, the coach is prompted to rewrite the original text by leveraging inappropriate knowledge or factual errors, thereby intentionally introducing misleading or incorrect information into the generated content. A detailed description of the pipeline is shown in Appendix H.

From these adversarial samples, we construct the preference dataset $\mathcal{D}_{\text{pref}} = \left\{\left(\mathcal{B}_i, q_{i,j}, \tilde{c}_{i,j}, c_{i,j}^{\text{neg}}\right)\right\}$ that captures hierarchical confusion scenarios encountered in medical practice. The KPO objective is defined as:

$$\mathcal{L}_{\text{KPO}} = -\mathbb{E}_{(\mathcal{B},q,\mathcal{X}^+,\mathcal{X}^-)\sim\mathcal{D}_{\text{pref}}}\left[\log \sigma\left(\beta\Delta\right)\right] \quad (10)$$

where $\Delta = \log \frac{P_\theta(\mathcal{X}^+|\mathcal{B},q)}{P_{\theta_{\text{ref}}}(\mathcal{X}^+|\mathcal{B},q)} - \log \frac{P_\theta(\mathcal{X}^-|\mathcal{B},q)}{P_{\theta_{\text{ref}}}(\mathcal{X}^-|\mathcal{B},q)}$, with θ_{ref} being the model after the first phase, which specifically enhances resistance to semantic similarity traps by amplifying differential between medically valid and deceptive representations.

Full-Chain Supervised Fine-Tuning Finally, building upon sub-question proficiency, the student model learns end-to-end medical reasoning. The training objective is:

$$\mathcal{L}_{SFT_{chain}} = -\sum_{i=1}^{N} \log P\left(\tilde{\mathcal{C}}_i \mid \mathcal{B}_i, Q_i; \theta\right) \quad (11)$$

which preserves the hierarchical structure of $\tilde{\mathcal{C}}_i$, enabling internalization of multi-step reasoning pathways for complex medical problem-solving.

4 Experiment

4.1 Experimental Setup

Data Curation We employ the open-source multiple-choice dataset m1k (Huang et al., 2025b)

as the distillation source (using the question datasets and discarding the answers from the original dataset), comprising a curated and strategically sampled medical corpus integrating questions from MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and HeadQA (Vilares and Gómez-Rodríguez, 2019). This dataset provides a robust foundation for subsequent experimental evaluations. For the teacher role, we utilize DeepSeek-R1-0528 (Guo et al., 2025), a state-of-the-art open-source reasoning model demonstrating performance parity with leading proprietary counterparts. For the coach role, we employ DeepSeek-V3-0324 (Liu et al., 2024), a general-purpose model that offers strong instruction-following capabilities with reduced inference overhead. This dual-model design enables efficient generation of high-quality chain-of-thought distillation data while retaining the reasoning fidelity required for medical refinement via our MedCoach methodology. Detailed information about the benchmarks is listed in Appendix G.

Model Training We employ Qwen2.5-7B-Instruct (Yang et al., 2024a) as the primary student model. Following the data refinement process, the student undergoes progressive distillation through three sequential phases: (1) sub-question supervised fine-tuning (SFT_{sub}), (2) knowledge-aware preference optimization (KPO), and (3) full-chain supervised fine-tuning (SFT_{chain}). This hierarchical approach implements curriculum-oriented knowledge transfer that gradually bridges the reasoning gap between the teacher and student. Detailed training configurations and hyperparameter settings are provided in Appendix B.

Baselines We evaluate our method against established distillation paradigms: SBS (Hsieh et al., 2023), featuring decoupled distillation of rationales and answers; Std-CoT (Magister et al., 2023), implementing standard CoT distillation via direct fine-tuning; MT-CoT (Li et al., 2022), employing multi-task optimization of answer prediction and reasoning chain generation; SCOTT (Wang et al., 2023), enhancing reasoning consistency through counterfactual training data integration. We also train several powerful medical reasoning language models with MedCoach (4B, 7B), and compare them with other advanced language models, encompassing reasoning and non-reasoning models, as well as medical-specialized and general-purpose models. Specifically, the general-purpose non-reasoning models include Qwen2.5-7B-Instruct and Qwen2.5-3B-Instruct (Yang et al., 2024a); the general-purpose reasoning models include Qwen3-4B (Yang et al., 2025); the medical non-reasoning models include UltraMedical-8B (v3, 3.1) (Zhang et al., 2024); and the medical reasoning models include HuatuoGPT-o1 (7B, 8B) (Chen et al., 2025b), MedReason-8B (Wu et al., 2025b), and m1-7b (1k, 23k) (Huang et al., 2025b).

Benchmarks and Metrics Following the work of Chen et al. (2025b), we primarily conduct evaluations on the test or validation sets of MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MMLU-Pro-Med (Wang et al., 2024b), and GPQA-Med (Rein et al., 2024), ensuring that these evaluation sets do not overlap with those used during the data curation phase. We use accuracy as the evaluation metric, which is also consistent with prior work. Furthermore, to enable a more comprehensive comparison of the robust large medical reasoning models we trained, and following Huang et al. (2025b), we perform more extensive benchmarking on additional diverse medical datasets (Lancet [1], MedBullets (Chen et al., 2025a), MedExpertQA (Zuo et al., 2025), NEJM [2]) to effectively demonstrate the models' strengths. Detailed information about the benchmarks is listed in Appendix G.

4.2 Overall Performance

We conduct extensive experiments across diverse medical benchmarks, with key results presented in Table 1. The empirical findings demonstrate that CoT prompting consistently enhances performance across all model architectures, including both lightweight student models and more sophisticated models with inherent reasoning capabilities. While most CoT distillation approaches show improved performance over baseline methods, they generally underperform standard CoT distillation. This phenomenon can be attributed to the inherent complexity and tight coupling of rationales generated by state-of-the-art reasoning models, which markedly differ from the constructed CoT data employed in previous studies. Our proposed method, MedCoach, achieves significant improvements over existing approaches, outperforming standard CoT distillation by 2.5% and surpassing baseline methods by 6.76%.

[1] https://www.thelancet.com/
[2] https://www.nejm.org/

Method	Distill?	MedMCQA	MedQA	PubMedQA	MMLU-Pro	GPQA	AVG
In-domain?		✓	✓	✓	×	×	
Teacher: Deepseek-R1-0528							
Baseline	-	80.20	93.07	68.32	90.10	73.27	80.99
CoT (Wei et al., 2022)	×	81.19	94.06	70.30	91.09	71.29	81.58
Coach: Deepseek-V3-0324							
Baseline	-	75.30	88.66	69.33	84.67	66.00	76.79
CoT (Wei et al., 2022)	×	74.26	89.11	74.26	87.13	66.34	78.22
Student: Qwen2.5-7B-Instruct							
Baseline	-	56.35	61.51	71.00	60.98	41.28	58.22
CoT (Wei et al., 2022)	×	58.32	62.60	68.40	64.00	44.40	59.54
MT-CoT (Li et al., 2022)	✓	57.16	67.87	72.30	64.56	47.18	61.81
SBS (Hsieh et al., 2023)	✓	57.45	69.05	72.80	61.89	43.85	61.01
SCOTT (Wang et al., 2023)	✓	51.64	61.35	71.30	55.31	39.49	55.82
Std-CoT (Magister et al., 2023)	✓	59.17	65.15	73.70	64.10	50.26	62.48
MedCoach w/ decom	✓	**60.80**	67.60	70.40	**71.80**	48.21	63.76
MedCoach w/o KPO	✓	60.40	71.80	74.40	64.60	50.51	64.34
MedCoach (Full Implement.)	✓	58.76	**73.66**	**75.30**	65.15	**52.05**	**64.98**

Table 1: Performance comparison of different methods across various medical QA benchmarks (Accuracy, %). "Baseline" denotes the direct test results of the model, while "CoT" represents leveraging CoT prompting.

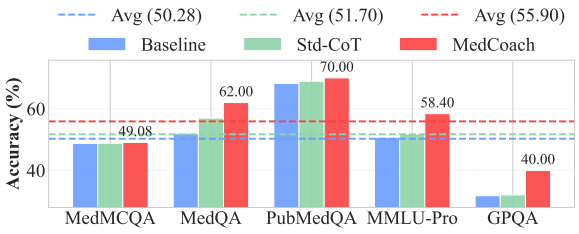

Figure 3: Performance on the weaker student model.

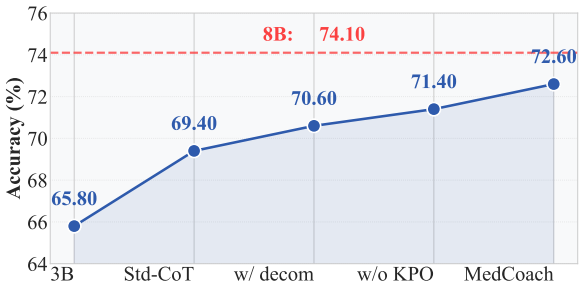

Figure 4: Performance when teacher and student models have relatively close parameter counts.

4.3 Ablation Study

Ablation on Modules We conduct a comprehensive ablation study on the three modules: question decomposition, knowledge grounding, and KPO. As shown in Table 1, each phase achieves improvements in effectiveness compared to the previous stage. On specific datasets, the method using only the question decomposition component (w/ decom, $SFT_{sub} + SFT_{chain}$ without rewriting and KPO) achieves better performance on MedMCQA and MMLU-Pro. This may be attributed to differences between the distribution of factual knowledge in each dataset and the domain specificity of the medical knowledge graph selected in this work. We further report an ablation study omitting chain-level training in Appendix C.

Strong Performance on Weaker Student Model We also conduct experiments on the relatively smaller model Qwen2.5-3B-Instruct, with results presented in Figure 3. It can be observed that our MedCoach method achieves strong performance across multiple medical QA benchmarks, improving by 4.2% over Std-CoT. In particular, it attains a score of 62.00% on MedQA, which is significantly superior to both the baseline and Std-CoT methods. This indicates that MedCoach helps bridge the capability gap between stronger and weaker models by providing a curriculum-oriented learning trajectory that smooths the steep reasoning curve.

Robust Performance over Relatively Close Models Although our framework was originally designed to alleviate the steep learning curve arising from the significant capability gap between strong and weak models, we further conduct experiments under a relatively small-gap setting. Specifically, we adopt a lightweight open-source reasoning model, Qwen3-8B, as both the teacher (serving

Model	MedMCQA	MedQA	PubMedQA	MMLU-Pro	GPQA	Lancet	MB_op4	MB_op5	MedX	NEJM	AVG-5	AVG
HuatuoGPT-o1-7B	63.73	72.11	78.20	66.78	46.67	64.32	56.17	52.92	14.91	65.67	65.50	58.15
HuatuoGPT-o1-8B	**63.78**	75.02	**80.50**	64.23	**57.18**	62.62	54.87	51.95	17.32	62.85	68.14	59.03
UltraMedical-8B-3	59.17	71.64	70.30	61.43	50.26	59.71	54.55	51.62	15.04	66.00	62.56	55.97
UltraMedical-8B-3.1	63.64	75.10	79.00	63.39	49.49	**67.23**	61.36	55.19	16.56	**66.67**	66.12	59.76
MedReason-8B	60.17	70.78	78.50	65.08	52.56	56.80	57.14	51.30	**18.29**	62.69	65.42	57.33
m1-7b-1k	57.85	69.05	76.90	61.63	44.10	60.68	57.47	51.62	15.04	60.20	61.91	55.45
m1-7b-23k	61.49	71.96	74.00	63.65	46.67	60.44	57.47	**57.79**	17.81	63.18	63.55	57.45
Qwen2.5-7B-Instruct	56.35	61.51	71.00	60.98	41.28	61.65	46.43	39.94	12.28	58.87	58.22	51.03
Meta-L3.1-8B-Instruct	56.13	61.27	77.70	57.79	39.49	58.98	50.65	45.13	15.04	58.37	58.48	52.06
Qwen3-4B	59.81	72.03	72.60	71.60	54.87	62.86	59.09	50.97	13.73	62.19	66.18	57.98
MedCoach-7B	58.76	73.66	75.30	65.15	52.05	57.52	58.44	52.27	14.63	59.54	64.98	56.73
MedCoach-4B	62.69	**78.65**	76.50	**71.79**	53.08	61.65	**61.69**	55.52	17.32	60.20	**68.54**	**59.91**

Table 2: Performance (accuracy in %) on medical QA benchmarks. Abbreviations: MMLU-Pro (Medical track), GPQA (Medical track), MB_op4/op5 (MedBullets Option 4/5), MedX (MedXpertQA), AVG-5: Average of MedMCQA, MedQA, PubMedQA, MMLU-Pro, GPQA; AVG: Overall average.

as the source of distillation data) and the coach (responsible for parsing difficulties and imparting knowledge), with Qwen2.5-3B-Instruct as the student model. All experiments under this setting are conducted exclusively on the PubMedQA benchmark, and the results (shown in Figure 4) indicate that our framework still delivers robust performance, with each module contributing incrementally, even enabling the student model to achieve performance very close to that of the teacher model.

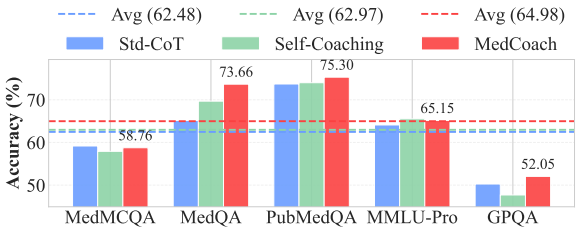

Figure 5: Performance of self-coaching.

Coaching by Student Itself In our framework, we employ a non-reasoning model (DeepSeek-V3-0324 (Liu et al., 2024)) as the coach primarily for its token-efficiency and instruction-following ability (Li et al., 2025c; Kwon et al., 2025), rather than using the more powerful teacher model (DeepSeek-R1-0528 (Guo et al., 2025)). To further investigate the coaching mechanism, we also experiment with a self-coaching setup in which the student model acts as its own coach. Results (see Figure 5) demonstrate that even a student model whose performance lags significantly behind the teacher can still achieve improvements when acting as its own coach. This highlights the effectiveness of the coaching paradigm in aiding problem-solving, paralleling how humans tackle complex questions through iterative refinement.

Comparison with Advanced Language Models Finally, we train a powerful medical reasoning model based on a strong and lightweight model, Qwen3-4B, using our innovative MedCoach distillation framework, and compare it with various open-source advanced models (see Table 2). Our MedCoach-7B outperforms UltraMedical-8B-3 and m1-7B-23k, which are trained on large-scale datasets with data volumes on the order of 1k, demonstrating strong performance. In particular, MedCoach-4B, with a very small parameter scale, surpasses HuatuoGPT-o1-8B and UltraMedical-8B-3.1, which are trained with more complex medical data, achieving state-of-the-art performance.

5 Conclusion

In conclusion, this work addresses key limitations in existing distillation methods for enhancing medical LLM reasoning under low-resource conditions. While LLMs like DeepSeek-R1 advance reasoning capabilities across domains, current distillation methods often neglect factual augmentation and the capability gap between teachers and students, hindering effective learning. We propose MedCoach, a novel framework that tackles these issues through three core components: question decomposition to smooth the learning curve, medical knowledge graph-based factual augmentation at the sub-question level to reduce hallucinations, and a knowledge perturbation strategy for finer-grained knowledge discrimination. These elements collectively enhance both knowledge grounding and reasoning ability.

Limitations and Future Work

We restrict our experiments to compact and mid-sized open models due to computational budget constraints; scaling the coaching paradigm to larger reasoning backbones (e.g., 14B–70B) and examining saturation or diminishing returns are deferred to future study. Our current KG integration adopts a lightweight semantic retrieval + rewrite strategy (single-hop, triple textualization, top-K filtering). While this already yields measurable factual gains, it does not yet exploit richer structured signals (e.g., multi-hop paths or hierarchical ontologies). The effectiveness of the framework is inherently coupled to the coverage and precision of the underlying knowledge graph; the current graph focuses primarily on precision medicine entities, and performance may vary when applied to subspecialties with sparser knowledge representation. While our results indicate improved performance, a more exhaustive mechanistic interpretation of how the coaching decomposition alters internal reasoning dynamics remains an open direction for future analysis.

Ethics Statements

All datasets used are public and contain no identifiable patient information. The model is developed strictly for research and must not be used for real-time clinical decision-making or patient-facing guidance. Although knowledge grounding and perturbation reduce hallucination risk, residual factual errors, bias, and coverage gaps may persist, especially for underrepresented conditions or populations.

Acknowledgements

This work was supported by the Guizhou Provincial Program on Commercialization of Scientific and Technological Achievements (Qiankehezhongyindi [2025] No. 006) and Alibaba Group through the Alibaba Innovation Research Program.

References

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. DISC-MedLLM: Bridging General Large Language Models and Real-World Medical Consultation. http://arxiv.org/abs/2308.14346.

Wenrui Cai, Chengyu Wang, Junbing Yan, Jun Huang, and Xiangzhong Fang. 2025. Enhancing reasoning abilities of small llms with cognitive alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7434–7449.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025b. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.

Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024. Learning to maximize mutual information for chain-of-thought distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6857–6868, Bangkok, Thailand. Association for Computational Linguistics.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small LLMs can master math reasoning with self-evolved deep thinking. In *Forty-second International Conference on Machine Learning*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

hongzhou yu, Tianhao Cheng, Yingwen Wang, Wen He, Qing Wang, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2025. FinemedLM-o1: Enhancing medical knowledge reasoning ability of LLM from supervised fine-tuning to test-time training. In *Second Conference on Language Modeling*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Xiang Huang, Jiayu Shen, Shanshan Huang, Sitao Cheng, Xiaxia Wang, and Yuzhong Qu. 2025a. TARGA: Targeted synthetic data generation for practical reasoning over structured data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2704–2726, Vienna, Austria. Association for Computational Linguistics.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025b. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*.

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

Runsong Jia, Mengjia Wu, Ying Ding, Jie Lu, and Yi Zhang. 2025. Hetgcot-rec: Heterogeneous graph-enhanced chain-of-thought llm reasoning for journal recommendation. *arXiv preprint arXiv:2501.01203*.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds ^{3}: Towards medical slow thinking with self-evolved soft dual-sided process supervision. *arXiv preprint arXiv:2501.12051*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

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 Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jerome P Kassirer. 2010. Teaching clinical reasoning: case-based and coached. *Academic medicine*, 85(7):1118–1124.

Yongchan Kwon, Shang Zhu, Federico Bianchi, Kaitlyn Zhou, and James Zou. 2025. Reasonif: Large reasoning models fail to follow instructions during reasoning. *arXiv preprint arXiv:2510.15211*.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, et al. 2025a. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*.

Feiyang Li, Peng Fang, Zhan Shi, Arijit Khan, Fang Wang, Dan Feng, Weihao Wang, Xin Zhang, and Yongjian Cui. 2025b. Cot-rag: Integrating chain of thought and retrieval-augmented generation to enhance reasoning in large language models. *arXiv preprint arXiv:2504.13534*.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025c. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *arXiv preprint arXiv:2505.11423*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching Small Language Models to Reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806.

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.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. *arXiv preprint arXiv:1906.04701*.

Guoxin Wang, Minyu Gao, Shuai Yang, Ya Zhang, Lizhi He, Liang Huang, Hanlin Xiao, Yexuan Zhang, Wanyue Li, Lu Chen, et al. 2025a. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*.

Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. 2024a. Boosting language models reasoning with chain-of-knowledge prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4958–4981, Bangkok, Thailand. Association for Computational Linguistics.

Lin Wang and Kuk-Jin Yoon. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*.

Teng Wang, Wing Yin Yu, Zhenqi He, Zehua Liu, HaileiGong HaileiGong, Han Wu, Xiongwei Han, Wei Shi, Ruifeng She, Fangzhou Zhu, and Tao Zhong. 2025b. BPP-search: Enhancing tree of thought reasoning for mathematical modeling problem solving. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–838, Vienna, Austria. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 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*.

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

Sibo Wei, Xueping Peng, Yifei Wang, Tao Shen, Jiasheng Si, Weiyu Zhang, Fa Zhu, Athanasios V Vasilakos, Wenpeng Lu, Xiaoming Wu, et al. 2025. Biancang: a traditional chinese medicine large language model. *IEEE Journal of Biomedical and Health Informatics*.

Dingjun Wu, Yukun Yan, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2025a. Kg-infused rag: Augmenting corpus-based rag with external knowledge graphs. *arXiv preprint arXiv:2506.09542*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. 2025b. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Yike Wu, Yi Huang, Nan Hu, Yuncheng Hua, Guilin Qi, Jiaoyan Chen, and Jeff Z. Pan. 2024. CoTKR: Chain-of-Thought Enhanced Knowledge Rewriting for Complex Knowledge Graph Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3501–3520, Miami, Florida, USA. Association for Computational Linguistics.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, et al. 2024b. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Advances in Neural Information Processing Systems*, 37:138632–138662.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan.

2024c. Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-World Multi-Turn Dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19368–19376.

Ken Yano, Zheheng Luo, Jimin Huang, Qianqian Xie, Masaki Asada, Chenhan Yuan, Kailai Yang, Makoto Miwa, Sophia Ananiadou, and Jun'ichi Tsujii. 2025. ELAINE-medLLM: Lightweight English Japanese Chinese Trilingual Large Language Model for Biomedical Domain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4670–4688, Abu Dhabi, UAE. Association for Computational Linguistics.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, et al. 2023. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. 2024. Distilling instruction-following abilities of large language models with task-aware curriculum planning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6030–6054.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Qi Zhao, Hongyu Yang, Qi Song, Xinwei Yao, and Xiangyang Li. 2025. Knowpath: Knowledge-enhanced reasoning via llm-generated inference paths over knowledge graphs. *arXiv preprint arXiv:2502.12029*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedxpertQA: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*.

A Statistic for Data

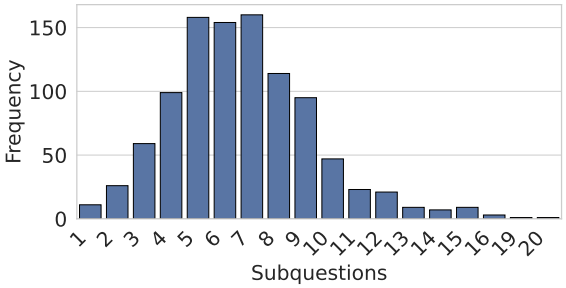

Figure 6: The statistics of sub-questions per response.

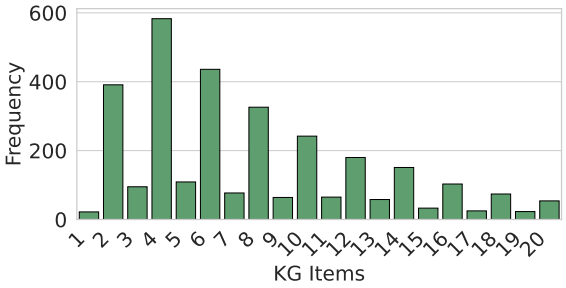

Figure 7: The statistics of relevant KG items per sub-question.

The original distillation question set contains 1k questions, with 9k decomposed sub-questions, among which 6k can be found in the corresponding reasoning text. The distribution of the number of sub-questions is shown in Figure 6. The distribution of KG triple entries corresponding to each sub-question is shown in Figure 7. The reason why the number of matched KG triple entries in the figure is mostly even is that the selected knowledge graph (PrimeKG (Chandak et al., 2023)) stores bidirectional edges, which are selected twice in the implementation. In terms of knowledge preference optimization, to avoid redundancy, at most two sub-questions are selected for each prompt for corresponding perturbation, yielding over 2k preference pair entries generated from almost 1k original sub-questions.

B Training Hyperparameters

For the primary 7B student model, the sub-question supervised fine-tuning (SFT_{sub}) phase employed a learning rate of 1×10^{-6} with a batch size of 128, warmup ratio of 0.1, and weight decay of 0.1. The knowledge-aware preference optimization (KPO) phase used a lower learning rate of 5×10^{-7} with a batch size of 16, warmup ratio of 0.1, weight

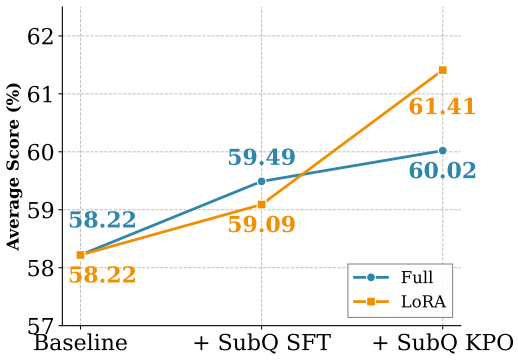

Figure 8: Performance of Sub-Question-Level Distillation Alone.

decay of 0.1, and $\beta = 0.1$. The final full-chain supervised fine-tuning (SFT_{chain}) phase adopted a learning rate of 1×10^{-5}, batch size of 16, warmup ratio of 0.05, and weight decay of 0.1. All training runs utilized the AdamW optimizer with a cosine learning rate scheduler.

C Effectiveness of Sub-Question-Level Distillation Alone

To isolate the contribution of fine-grained knowledge grounding, we evaluate a variant that omits the final chain-level supervised fine-tuning stage, applying only the sub-question-level supervised fine-tuning and preference optimization steps. Experiments are conducted on Qwen2.5-7B-Instruct under both full-parameter and LoRA settings, using the same five benchmarks summarized in Table 1. As shown in Figure 8, both configurations yield consistent improvements over the baseline. This demonstrates that fine-grained alignment on sub-questions alone effectively transfers the medical knowledge embedded in the teacher's reasoning chains and knowledge graph, achieving clear improvements even without chain-level training. The finding underscores the standalone contribution of the decomposition and augmentation stages. Notably, LoRA method attains a higher final score than full fine-tuning in the last stage, suggesting that its constrained parameter updates may enhance sensitivity to the subtle knowledge perturbations introduced during preference optimization.

D The Performance of Direct Rewriting

For knowledge rewriting, if sub-questions are not decomposed first, training on directly rewritten data yields very poor results, as shown in Table 3. This is because it is difficult for the overall CoT to rely solely on the embedding model to find corresponding similar knowledge, and the inherent simplification in the rewriting task can severely impair the model's reflective ability.

Dataset	Baseline	Std-CoT	DR
MedMCQA	56.49	59.17	54.36
MedQA	61.43	65.15	61.27
PubMedQA	71.00	73.70	75.10
MMLU-Pro	61.69	64.10	58.76
GPQA	40.77	50.26	46.41
AVG	58.28	62.48	59.18

Table 3: Disastrous performance by direct rewriting (Accuracy, %). DR denotes the direct rewriting method.

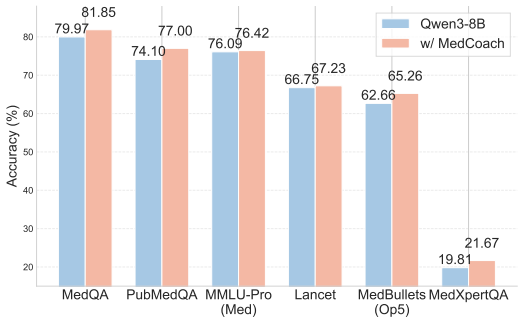

Figure 9: Accuracy comparison between the baseline Qwen3-8B model and its MedCoach-distilled variant.

E Generalization to Stronger Backbone Models

To examine whether MedCoach provides additional benefits when the student model already possesses strong intrinsic reasoning capabilities, we conduct experiments using Qwen3-8B as the backbone. As shown in Figure 9, Qwen3-8B achieves competitive performance across several benchmarks. Applying our proposed curriculum distillation framework results in modest but consistent improvements.

F Description and Textualization of PrimeKG

PrimeKG (Precision Medicine Knowledge Graph) is a multimodal knowledge graph designed for precision medicine analyses. The graph integrates 20 standard biomedical resources , comprising 129,375 nodes and 4,050,249 relationships to describe 17,080 unified disease entities. Structurally, PrimeKG captures information across ten major biological scales, including disease-associated protein perturbations, biological processes and path-

ways, anatomical regions, clinical phenotypes, environmental exposures, and approved drugs along with their therapeutic targets. Regarding node associations, in addition to standard indications, the graph systematically incorporates an abundance of contraindications and off-label use edges to support computational analyses of drug-disease networks. Furthermore, demonstrating its multimodal nature, PrimeKG supplements its traditional structural network with natural language text descriptions extracted from clinical guidelines and knowledge bases for both drug and disease nodes. For disease entity processing, the project utilized natural language processing techniques to collapse and unify overlapping disease concepts across various ontologies (such as UMLS, MONDO, and Orphanet) to improve the correspondence between disease nodes in the graph and actual clinical subtypes.

There are generally two ways to textualize the triples of large-scale knowledge graphs: direct format-matching textualization and textualization using a smaller language model. Direct textualization tends to lose significant semantic information. Although using a smaller language model can provide a better understanding of knowledge triples, its grasp of specialized medical knowledge remains limited. We observed that, for the knowledge graph (PrimeKG (Chandak et al., 2023)) we use, the combinations of entity types and relationship types are very limited, as shown in Table 4. This indicates that the text descriptions corresponding to the triples are highly similar. Therefore, we employ the powerful open-source reasoning model Deepseek-R1-0528 (Guo et al., 2025) to obtain the textualization format corresponding to each specific combination.

G Detailed Information of Datasets and Benchmarks

We use only the question set of the m1k dataset (Huang et al., 2025b) as the distillation source, excluding its corresponding answers. Detailed information about the datasets and benchmarks involved in training and evaluation is provided as follows:

- **m1k** (Huang et al., 2025b): The m1k dataset is a high-quality medical QA collection derived from 196K publicly available samples across four sources: MedMCQA, MedQA-USMLE, HeadQA, and PubMedQA. Through a rigorous refinement process, the dataset was filtered to 37K challenging questions that neither Qwen2.5-7B nor Qwen2.5-32B could solve correctly. DeepSeek-R1 then generated verified CoT solutions, yielding 23K correct reasoning paths. A final diversity sampling step produced the 1K core training set (m1k), balanced across medical domains (via MeSH categories) and source datasets.

- **MedMCQA** (Pal et al., 2022): A large-scale multiple-choice medical question-answering dataset compiled from authentic exam questions from India's All India Institute of Medical Sciences (AIIMS) and the National Eligibility cum Entrance Test for Postgraduate (NEET-PG). It includes over 194,000 high-quality medical questions spanning 2,400 health topics across 21 medical disciplines, demonstrating substantial topic diversity. We use 4,183 questions as the test set, following (Chen et al., 2025b). It is distributed under Apache-2.0 license.

- **MedQA** (Jin et al., 2021): A USMLE-based clinical medicine question bank comprising 12,723 questions derived from 18 commonly used authoritative clinical medicine textbooks. These questions cover a diverse range of clinical medicine subjects and require professional-level reasoning through integration of multi-source evidence. We use 1,273 questions as the test set, following (Chen et al., 2025b). It is distributed under the CC-BY-4.0 license.

- **PubMedQA** (Jin et al., 2019): A biomedical question-answering dataset derived from PubMed abstracts, featuring 1,000 multiple-choice question-answering examples (with options yes/no/maybe) annotated by experts. The dataset's knowledge foundation draws from a vast collection of 211,300 PubMed articles. The core task is to determine the correct answer to each research question by analyzing and interpreting the content within the provided abstracts. The test set contains 1,000 questions. It is distributed under MIT license.

- **HeadQA** (Vilares and Gómez-Rodríguez, 2019): A multiple-choice dataset focused on healthcare. Its questions are derived from examinations intended to qualify individuals for specialized roles within Spain's health-

Relation Type	Relation Name	Entity Type I	Entity Type II
anatomy_anatomy	parent-child	anatomy	anatomy
anatomy_protein_absent	expression absent	anatomy	gene/protein
anatomy_protein_present	expression present	anatomy	gene/protein
bioprocess_bioprocess	parent-child	biological_process	biological_process
bioprocess_protein	interacts with	biological_process	gene/protein
cellcomp_cellcomp	parent-child	cellular_component	cellular_component
cellcomp_protein	interacts with	cellular_component	gene/protein
contraindication	contraindication	disease	drug
indication	indication	disease	drug
off-label use	off-label use	disease	drug
disease_disease	parent-child	disease	disease
disease_phenotype_negative	phenotype absent	disease	effect/phenotype
disease_phenotype_positive	phenotype present	disease	effect/phenotype
disease_protein	associated with	disease	gene/protein
drug_drug	synergistic interaction	drug	drug
drug_effect	side effect	drug	effect/phenotype
drug_protein	carrier	drug	gene/protein
drug_protein	enzyme	drug	gene/protein
drug_protein	target	drug	gene/protein
drug_protein	transporter	drug	gene/protein
exposure_bioprocess	interacts with	biological_process	exposure
exposure_cellcomp	interacts with	cellular_component	exposure
exposure_disease	linked to	disease	exposure
exposure_exposure	parent-child	exposure	exposure
exposure_molfunc	interacts with	exposure	molecular_function
exposure_protein	interacts with	exposure	gene/protein
molfunc_molfunc	parent-child	molecular_function	molecular_function
molfunc_protein	interacts with	gene/protein	molecular_function
pathway_pathway	parent-child	pathway	pathway
pathway_protein	interacts with	gene/protein	pathway
phenotype_phenotype	parent-child	effect/phenotype	effect/phenotype
phenotype_protein	associated with	effect/phenotype	gene/protein
protein_protein	ppi	gene/protein	gene/protein

Table 4: The combinations of entity and relation types in PrimeKG.

Label	Content
question	Which of the following acts as intermediate host of malaria parasite? A. Human B. Female anopheles C. Thromboculid mite D. Culex
sub-question	Are Thromboculid mites involved in the transmission of malaria?
original text	I know that chiggers are involved in scrub typhus, which is caused by Orientia tsutsugamushi, and the vector is a trombiculid mite. But that's not related to malaria. Malaria is not transmitted by mites.
chosen	No, Thromboculid mites are not involved in the transmission of malaria. While these mites are known vectors for scrub typhus (caused by Orientia tsutsugamushi), malaria is a distinct vector-borne disease transmitted exclusively by Anopheles mosquitoes. The two conditions, though both vector-borne, are unrelated in terms of their transmission mechanisms.
rejected	Yes, Thromboculid mites transmit malaria, just as they are involved in the transmission of epidemic louse-borne typhus and other vector-borne diseases.

Table 5: A case study.

care sector, posing significant challenges even to highly trained professionals. These questions are crafted by the Ministry of Health, Consumer Affairs and Social Welfare. This dataset is part of m1k and was not used directly. It is distributed under MIT license.

- **MMLU-Pro (Medical)** (Wang et al., 2024b): A specialized subset of the Massive Multitask Language Understanding (MMLU) benchmark focusing on professional medicine and related disciplines. Our evaluation centers on the Professional Medicine section. The dataset splits adhere to the methodology outlined in (Chen et al., 2025b). It is distributed under MIT license.

- **GPQA (Medical)** (Rein et al., 2024): The biomedical subset of the Graduate-Level QA dataset, featuring highly challenging, expert-designed multiple-choice questions that demand advanced reasoning. These questions are intentionally designed to be resistant to simple search-engine lookups. We evaluate on the biology/medical subset and follow the same splits as (Chen et al., 2025b). It is distributed under the CC-BY-4.0 license.

- **Lancet & NEJM**: Two small-scale QA datasets curated from clinical case reports published in *The Lancet* and the *New England Journal of Medicine* (NEJM), curated by Huang et al. (2025b).

- **MedBullets** (Chen et al., 2025a): A set of practice questions sourced from the MedBullets medical education platform. Following (Huang et al., 2025b), we use subsets classified as levels 4 and 5 (on a 1–5 scale, with 5 being the hardest), referred to as MedBullets_Op4 and MedBullets_Op5, each containing approximately 100 questions. These serve as rigorous benchmarks for model performance.

- **MedXpertQA** (Zuo et al., 2025): A dataset of 50 expert-authored, multi-step medical reasoning questions designed for in-depth qualitative evaluation. It is distributed under MIT license.

H Case Study

As shown in Table 5, when presented with a multiple-choice question like "Which acts as the intermediate host of malaria parasite?", we decompose it into sub-questions such as "Are Thromboculid mites involved in malaria transmission?" This decomposition is appropriate because it targets a less obvious option (C) that could confuse students.

By isolating this sub-question, we directly verify whether mites play any role in malaria's lifecycle, thereby streamlining the elimination process. Solving this sub-question aids the original question by conclusively excluding option C, narrowing choices to the scientifically valid options (A/B/D).

The initial unedited reply ("I know chiggers transmit scrub typhus...") had minor issues: while factually correct, its informal tone and lack of explicit negation (e.g., "no" or "incorrect") introduced slight uncertainty about its intent to disqualify option C. The rewritten response ("No, Thromboculid mites are not involved...") correctly incorporated additional knowledge by: (1) explicitly denying mites' role in malaria, (2) contrasting their confirmed role in scrub typhus, and (3) emphasizing malaria's exclusive dependence on Anopheles mosquitoes. This version is accurate because it aligns with established parasitology (WHO/CDC guidelines) and eliminates ambiguity through structured logic.

For the sub-question asking whether Thromboculid mites transmit malaria, the chosen response is superior because it accurately clarifies that these mites are unrelated to malaria transmission (a disease exclusively spread by Anopheles mosquitoes) while correctly associating them with scrub typhus, a distinction critical for scientific precision. It avoids misinformation by explicitly differentiating vector mechanisms, thus providing educational value. In contrast, the rejected response is problematic as it incorrectly asserts Thromboculid mites as malaria vectors, conflating them with typhus transmission, a major factual error that could propagate dangerous misunderstandings about disease prevention. The chosen answer's adherence to evidence-based parasitology makes it a reliable reference, whereas the rejected one exemplifies a high-risk inaccuracy requiring correction in training data.

I Prompt Templates

Each function represents a prompt applied to the LLM, which directly instructs the LLM to perform its respective task, aiming to facilitate the application of complex functions. Specific descriptions of each prompt are presented in the corresponding figures.

The Prompt Template for Context Extraction

Extract the background context from the following medical question. The context should include only factual statements or setup information, not the question itself. If there is no such context, return an empty string for "context".

Example:
Original Question:
"An otherwise healthy 30-year-old woman experiences intermittent headaches for 2 weeks. What is the likely diagnosis?"
Response:
{{"context": "An otherwise healthy 30-year-old woman experiences intermittent headaches for 2 weeks."}}

Now process the question below:
Original Question:
`{prompt}`

Return ONLY a JSON object with a single key "context", for example:
{{"context": "..."}}

Figure 10: Template for Context Extraction. Placeholders `{prompt}` denote the input of Multiple-choice Questions (MCQs).

The Prompt Template for Question Decomposition

You are an expert in medical question decomposition.
Using the reasoning process below, break the original complex question into a sequence of self-contained sub-questions.
- Each sub-question should correspond to a distinct step or critical content in the reasoning process.

- Pay special attention to any points of uncertainty or where deeper medical knowledge was invoked.

Reasoning Process:
`{think}`

Original Question:
`{prompt}`

Return ONLY a JSON array of strings, e.g.: ["First self-contained sub-question?", "Second self-contained sub-question?", ".."]

Figure 11: Template for question Decomposition. Placeholders `{think}` denote the reasoning content. Placeholders `{prompt}` denote the original question.

The Prompt Template for Segment Alignment

You are an information-retrieval specialist. Given the full reasoning process below and a specific sub-question, locate the single complete sentence or paragraph from the reasoning that answer or are related to the sub-question.
- Preserve the text exactly as it appears: do NOT paraphrase, shorten, or modify punctuation or capitalization.
- Return ONLY a JSON object with exactly two fields:
1. "subquestion": the original sub-question string
2. "grounded_text": the exact sentence or paragraph from the reasoning

Reasoning Process:
`{think}`

Sub-question:
`{subq}`

Example output:
{"subquestion": "subq", "grounded_text": "<exact sentence or paragraph>"}}

Figure 12: Template for Segment Alignment. Placeholders `{think}` denote the reasoning content. Placeholders `{subq}` denote the sub-question.

The Prompt Template for Relevance Judgment

You are a knowledge relevance classifier. Given a question, its corresponding answer text, and a list of knowledge triples, select only those triples that are relevant for rewriting the text.

Question:
`{subq}`
Corresponding Text:
`{gt}`
Knowledge Triples (one per line):
`{relevant_kg}`
Return ONLY a JSON object with a single field:
{"relevant_kg": [<the relevant triples exactly as in the input>]}
If none are relevant, return {"relevant_kg": []}.
Do not output anything else.

Figure 13: Template for Relevance Judgment. Placeholders `{subq}` denote the sub-qeustion. Placeholders `{gt}` denote the sub-solution. Placeholders `{relevant_kg}` denote the relevant knowledge triples.

The Prompt Template for Knowledge Grounding

You are a medical text rewriting assistant. Your task is to produce a clear, coherent answer that DIRECTLY addresses the given question by integrating relevant knowledge triples.

Context:
`{context}`
Question:
`{subq}`
Original text:
`{gt}`
Knowledge Triples (one per line):
`{kg_block}`

Note: All relationships are undirected (e.g., "parent-child" indicates a connection without specifying direction.
Please rewrite the original text into a com-

plete, self-contained answer to the question. You may modify, expand, or reorganize the original text, and flexibly apply the knowledge triples.
When using the provided knowledge triples, only incorporate those that are truly relevant and helpful for answering the question. Ignore any knowledge that is not directly applicable to solving the problem at hand.
If the original text contains factual errors or excessive uncertainty that would prevent a clear answer to the question, you may correct these issues and change the meaning as needed to better address the given question.
If the given knowledge triples expression is slightly inaccurate or broad, you should apply the corresponding correct version.
Return ONLY a JSON object with the single field:
{"rewritten_text": "<your rewritten answer >"}
Do NOT include any additional explanation or mention retrieval.

Figure 14: Template for Knowledge Grounding. Placeholders `{context}` denote the context information. Placeholders `{subq}` denote the sub-question. Placeholders `{gt}` denote the sub-solution. Placeholders `{kg_block}` denote the relevant knowledge.

The Prompt Template for Irrelevance Perturbation

Here is an one-shot example:
Question: What is the role of hemoglobin?
Correct answer: Hemoglobin carries oxygen in red blood cells.
Irrelevant triple: Hemoglobin functions as a digestive enzyme in the stomach.
Wrong answer: Hemoglobin functions to break down proteins during digestion.

Now generate a WRONG answer for the new data using IRRELEVANT triples:
Context:
`{context}`
Question:
`{subq}`
Correct answer:
`{a}`
Irrelevant Triples:
`{block}`

Return ONLY a JSON object with exactly one key "wrong_answer", e.g.:
{{"wrong_answer":"..."}}

Figure 15: Template for Irrelevance Perturbation. Placeholders `{context}` denote the context information. Placeholders `{subq}` denote the sub-question. Placeholders `{a}` denote the knowledge-grounded sub-solution. Placeholders `{block}` denote the irrelevant knowledge triple.

The Prompt Template for Swap Perturbation

Here is an one-shot example:
Question: How does acetylcholine work in the synapse?
Correct answer: It binds nicotinic and muscarinic receptors.
Swapped triple: Acetylcholine transports oxygen in red blood cells.
Wrong answer: Acetylcholine carries oxygen in red blood cells.

Now generate a WRONG answer for the new data using SWAPPED triples:
Context:
`{context}`
Question:
`{subq}`
Correct answer:
`{a}`
Swapped Triples:
`{block}`

Return ONLY a JSON object with exactly one key "wrong_answer", e.g.:
{{"wrong_answer":"..."}}

Figure 16: Template for Swap Perturbation. Placeholders `{context}` denote the context information. Placeholders `{subq}` denote the sub-question. Placeholders `{a}` denote the knowledge-grounded sub-solution. Placeholders `{block}` denote the swapped knowledge triple.

The Prompt Template for Negation Perturbation

Here is an one-shot example:
Question: What does cortisol do?
Correct answer: It helps the body respond to stress.
Negated triple: Cortisol does not regulate stress response.
Wrong answer: Cortisol has no role in stress response.

Now generate a WRONG answer for the new data using NEGATED triples:
Context:
`{context}`
Question:
`{subq}`
Correct answer:
`{a}`
Negated Triples:
`{block}`

Return ONLY a JSON object with exactly one key “wrong_answer", e.g.:
{{“wrong_answer":“..."}}

Figure 17: Template for Negation Perturbation. Placeholders `{context}` denote the context information. Placeholders `{subq}` denote the sub-question. Placeholders `{a}` denote the knowledge-grounded sub-solution. Placeholders `{block}` denote the negated knowledge triple.