Med-R^3: Enhancing Medical Retrieval-Augmented Reasoning of LLMs via Progressive Reinforcement Learning

Anonymous ACL submission

Abstract

In medical scenarios, effectively retrieving external knowledge and leveraging it for rigorous logical reasoning is of significant importance. Despite their potential, existing work has predominantly focused on enhancing either retrieval or reasoning capabilities of the models in isolation, with little attention given to their joint optimization, which leads to limited coordination between the two processes. Additionally, current methods rely heavily on supervised fine-tuning (SFT), which can cause models to memorize existing problem-solving pathways, thereby restricting their generalization ability when confronted with novel problem contexts. Furthermore, while some studies have explored to improve retrieval-augmented reasoning in general domains via reinforcement learning, their reward function designs do not adequately capture the specific demands of the medical domain. To address these challenges, we introduce **Med-R^3**, a **Med**ical **R**etrieval-augmented **R**easoning framework driven by progressive **R**einforcement learning. In this framework, we first develop the model's ability to perform logical reasoning over medical problems. Subsequently, on the basis of this foundation, we adaptively optimize the retrieval capability to better align with the characteristics of knowledge corpus and external information utilization throughout the reasoning process. Finally, we conduct joint optimization of the model's retrieval and reasoning coordination. Extensive experiments indicate that **Med-R^3** could achieve state-of-the-art performances, with LLaMA3.1-8B-Instruct + Med-R^3 surpassing closed-sourced GPT-4o-mini by 3.93% at a comparable parameter scale, while Qwen2.5-14B augmented with Med-R^3 shows a more substantial gain of 13.53%.

1 Introduction

With the rapid development of artificial intelligence, large language models (LLMs) have shown remarkable potential in various fields (Lewkowycz et al., 2022). However, when applied to the medical domain, LLMs face unique challenges. Accurate diagnosis of medical conditions requires rigorous logical reasoning (Lucas et al., 2024; Savage et al., 2024), yet unlike domains such as mathematics or programming, where internal knowledge is often sufficient, the complexity and specificity of medical diagnosis necessitate the integration of external, up-to-date, and domain-specific knowledge (Xiong et al., 2024a; Lu et al., 2025). Therefore, during the process of solving medical problems, both *retrieval* and *reasoning* play crucial roles.

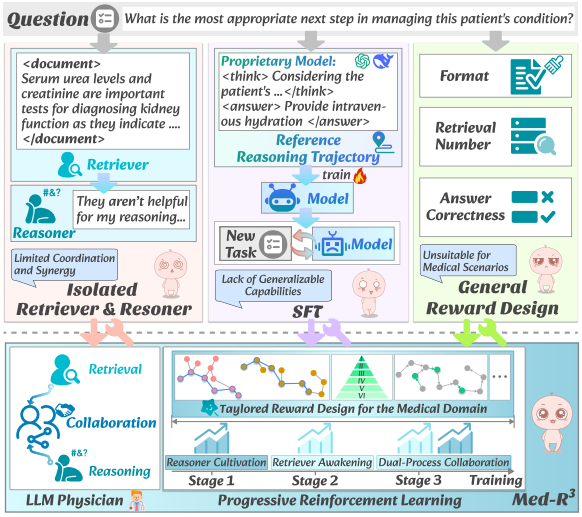

Figure 1: Comparison of Med-R^3 (*bottom*) with existing methods (*top*) for medical problem-solving.

Despite their potential, existing works face some limitations, as depicted in Figure 1: ***(C1) Limited Coordination between Retrieval and Reasoning***: While research has been devoted to independently improving the retrieval (Jeong et al., 2024; Xiong et al., 2024a,b) and reasoning (Goh et al., 2024; Lucas et al., 2024) capabilities of models in the medical domain, they do not adequately align the retrieval strategy with requirements of the reasoning process, which is a critical gap that limits the system's end-to-end performance (Yoran et al.,

2023; Asai et al., 2024). ***(C2) Lack of Generalization in SFT***: Subsequent studies have investigated to achieve the joint optimization of these components during the supervised fine-tuning (SFT) stage (Wang et al., 2024). However, recent findings suggest that SFT inherently causes models to memorize task-specific shortcuts rather than learning generalizable reasoning in novel scenarios (Lai et al., 2025). Conversely, reinforcement learning (RL) has shown high effectiveness at enhancing sophisticated reasoning capabilities in LLMs (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025). ***(C3) Tailored Reward Design for Medical Scenarios***: While recent works such as R1-Searcher (Song et al., 2025) and ReSearch (Chen et al., 2025) have explored training models for retrieval-augmented reasoning via reinforcement learning in general-domain tasks, their reward strategies are not well suited to the medical domain, where factors including the coverage of entities and relationships, as well as the credibility of retrieved documents, are of paramount importance in the reasoning process.

To address these challenges, we introduce ***Med-R***^{3}, a ***Med***ical ***R***etrieval-augmented ***R***easoning framework driven by progressive ***R***einforcement learning. For ***C1*** and ***C2***, we perform a progressive RL to co-optimize the model's retrieval and reasoning capabilities in three distinct stages: **(1) Stage 1: Reasoner Cultivation.** We begin by developing the model's logical reasoning abilities when addressing medical questions. **(2) Stage 2: Retriever Awakening.** Building upon the reasoning capabilities acquired in Stage 1, we adaptively optimize the retriever to better align with the retrieval features of the knowledge base and external information utilization during the model's reasoning process. **(3) Stage 3: Joint Optimization.** Finally, we refine the coordination between retrieval and reasoning in models to enhance their collaborative performance in medical scenarios. For ***C3***, we design specialized rewards tailored to the characteristics of medical reasoning at each training stage, encompassing aspects such as the coverage of entities and relations during the reasoning process, as well as the effectiveness of retrieval mechanisms that jointly consider the quality of medical evidence (Sackett et al., 1996) and the influence of retrieved documents within the overall reasoning trajectory. The main contributions can be summarized as follows:

- *Training Framework Advancement.* We propose Med-R^3, a novel training framework to improve the retrieval-augmented reasoning performance within the medical scenario based on progressive reinforcement learning, which jointly enhances the model's capability in retrieval and reasoning through a three-stage co-optimization strategy.

- *Reward Design Innovation.* Considering the unique characteristics of logical inference during medical problem-solving, we design reward metrics specifically tailored to the medical domain to supervise the reasoning process.

- *Performance and Effectiveness.* Experiments indicate that models trained with Med-R^3 significantly improve medical performances. Notably, LLaMA3.1-8B-Instruct + Med-R^3 surpasses the closed-sourced proprietary model GPT-4o-mini by 3.93% at a comparable parameter scale, while Qwen2.5-14B integrated with Med-R^3 shows a more substantial gain of 13.53%.

2 Med-R^3

2.1 Training Data Construction

Inspired by the data selection and verifiability transformation process of HuatuoGPT-o1 (Chen et al., 2024b), we filtered medical questions based on their cognitive complexity and compatibility with reinforcement learning, and then constructed reference reasoning trajectories for verifying.

2.1.1 Data Selection

We utilize both closed-set exam questions and rare disease diagnostic datasets for our progressive reinforcement learning (RL) phase, including the training sets from MedQA-USMLE, MedQA-MCMLE (Jin et al., 2020), MedMCQA (Pal et al., 2022), and RareArena[1] (THUMedInfo, 2025). Data filtering and processing are conducted referring to the following criteria:

- *Content Suitability Judgment*: We employed DeepSeek-V3 (Liu et al., 2024) to exclude the medical questions that are unsuitable for RL or lacking unique answers (e.g., those requiring the identification of incorrect options).

- *Reasoning Complexity Filtering*: We then used pass@n metrics to assess the question difficulty. Specifically, we employed DeepSeek-R1 (Guo et al., 2025) to generate

[1]For the RareArena dataset, we split the data into training and test sets with a ratio of 8:2 through random sampling.

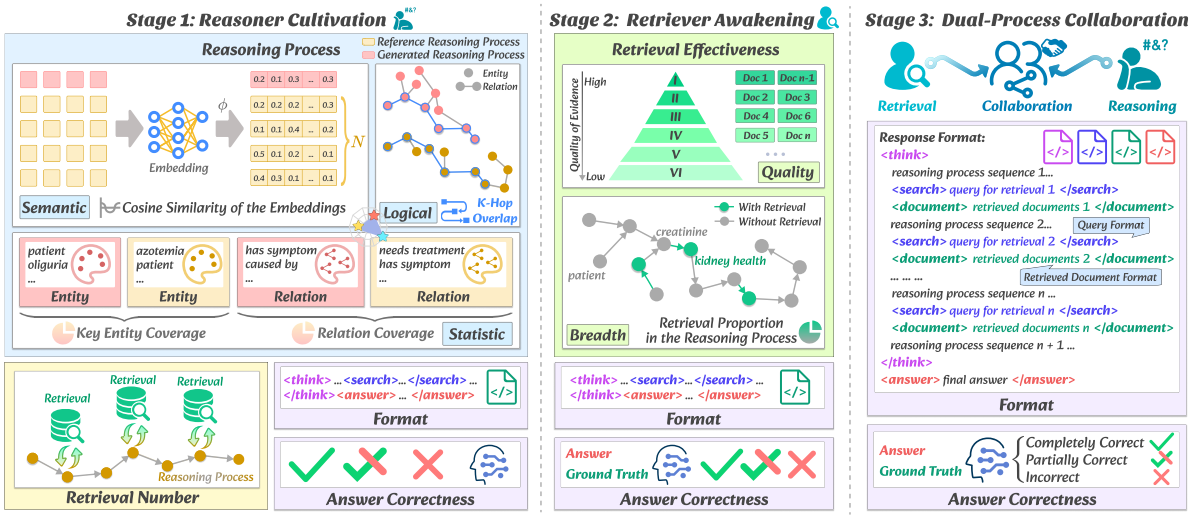

Figure 2: The progressive reinforcement learning pipeline of ***Med-R***^{3}. We designed tailored rewards for different training stages to facilitate the effective optimization of the model's ability to interpret and answer medical questions.

up to $n = 16$ answers along with the corresponding reasoning process for each medical question. DeepSeek-V3 was then utilized to evaluate both the intermediate reasoning steps and final answers, with scores ranging from 1 to 5. We filtered out questions that were either too simple (achieving full correctness within a few rollouts, $n' \leq 3$) or excessively ambiguous (consistently receiving scores of 2 or lower across all 16 rollouts). Finally, we obtained a training dataset comprising 2,140 questions from MedQA-USMLE, 1,204 from MedQA-MCMLE, 6,748 from MedMCQA, and 429 from the RareArena dataset.

2.1.2 Construction of Reference Reasoning Trajectories

We performed data augmentation by rephrasing closed-set questions into open-ended formats, generating high-quality reasoning process, and further extracting structured reasoning paths that capture the pivotal logical steps.

- *Open-Ended Standardization*: We then reformatted the multiple-choice questions into open-ended formats using DeepSeek-V3 (Liu et al., 2024), transforming questions with predefined answer choices into tasks that require free-form, reasoning-based responses.

- *Reasoning Process Generation*: Similar to the rollout process of data selection based on reasoning complexity, we prompted DeepSeek-R1 to generate all possible reasoning paths along with final answers for each medical question. During the reasoning process, the model dynamically retrieves relevant external knowledge as needed and integrates the retrieved information to support continued inference. Then we employed DeepSeek-V3 as an evaluator to assess the generated reasoning processes and answers on a scale from 1 to 5, retaining those with a score of 5. This procedure yielded 5 high-quality reasoning processes per question on average.

- *Medical Knowledge Graph Extraction*: Different from mathematical problems, which typically rely on rigid sequential logic, medical reasoning demands an understanding of rich, multi-relational structures that link symptoms, diagnoses, and therapeutic interventions. To better capture such a relational structure, we formalize reasoning processes as medical knowledge graph triplets, enabling a more precise comparison of coverage in both entities (e.g., disease, surgery, etc.) and relations. Here we utilize DeepSeek-V3 to extract the medical knowledge graph from the natural language reasoning process:

$$\mathcal{G} = \{(h, l, t, s) \mid h \in \mathcal{E}, l \in \mathcal{L}, t \in \mathcal{E}, s \in \mathbb{B}\} \quad (1)$$

where $\mathcal{E}$ denotes the set of all entities, $\mathcal{L}$ represents the set of all relations, and (h, l, t) is a triple in which the head entity h is connected to the tail entity t via the relation l. For each triplet, we add a binary attribute s, where $\mathbb{B} = \{0, 1\}$, specifying whether it was obtained through external knowledge retrieval

($s = 1$) or internally generated via the model's reasoning capabilities ($s = 0$).

2.2 Progressive Reinforcement Learning

We employ a three-stage pipeline to enhance the model's capabilities in retrieval and reasoning for medical problems, as illustrated in Figure 2.

2.2.1 Stage 1: Cultivating the Reasoner

In this stage, we focus on developing the model's ability to reason through medical problems. Unlike previous works (Chen et al., 2024b; Liu et al., 2025a) that typically verify model performance based only on the correctness of the final answer, we incorporate a reasoning process evaluation mechanism into the reward function, aligning closely with the characteristics of clinical decision-making. Overall, the reward function design in Stage 1 is *the sum of normalized* components:

Format. The model is required to produce its outputs according to a predefined output paradigm. Specifically: ***(1)*** All responses must strictly adhere to the format of `<think>...</think><answer>...</answer>`, and no duplicate tags are allowed within the response. ***(2)*** When retrieving the external knowledge is necessary, the corresponding query should be delineated using the `<search>...</search>` tag pair, and then the retrieved documents from knowledge corpus are encapsulated within `<document>...</document>` tags. Based on the above requirements, the reward is defined as:

$$\mathcal{R}_{format} = \begin{cases} 1, \text{if the format is correct} \\ 0, \text{if the format is incorrect} \end{cases} \tag{2}$$

Answer Correctness. In contrast to mathematical problems, medical questions often lack a strictly defined ground truth, as models may express the correct response with synonyms or paraphrases. As a result, exact string matching is *not* an effective method for verifying correctness in medical tasks. To address this, we employ a frontier model (e.g., DeepSeek-V3) as the evaluator to score the generated answers. It assesses whether the model's output semantically aligns with the reference ground truth. Answers are assigned a score of 2 for fully correct, 1 for partially correct (e.g., providing multiple answers, with at least one matching the correct answer), and 0 for incorrect responses:

$$\mathcal{R}_{answer} = \begin{cases} 2, \text{if completely correct} \\ 1, \text{if partially correct} \\ 0, \text{if incorrect} \end{cases} \tag{3}$$

Reasoning Process. This aspect constitutes a core component in fostering the model's deliberative reasoning capabilities during Stage 1. Given the characteristics of medical reasoning, which requires a holistic understanding of interconnected entities and their relationships—such as those between symptoms, diagnoses, and therapeutic interventions (Wu et al., 2023; Gao et al., 2025), we design the following reward for evaluation:

$$\mathcal{R}_{reasoning} = \tilde{R}_{semantic} + \tilde{R}_{statistic} + \tilde{R}_{logical} \tag{4}$$

where $\tilde{R}$ is the normalized score of R. Given N reference reasoning processes $\left\{r_{ref}^{(i)}\right\}_{i=1}^{N}$, $R_{semantic}$, $R_{statistic}$ and $R_{logical}$ of the current generated response r_{gen} are:

- $R_{semantic}$: This aspect measures the semantic alignment between the reasoning process generated by the model and those previously constructed by the frontier model for reference, which is quantified by the cosine similarity (Salton et al., 1975). $\phi(r)$ denotes the embedding of the reasoning process utilized the embedding model (Chen et al., 2024a).

$$R_{semantic} = \max_{i \in [1,N]} cos\Big(\phi\big(r_{\text{ref}}^{(i)}\big),\ \phi\big(r_{\text{gen}}\big)\Big) \tag{5}$$

- $R_{statistic}$: Medical reasoning typically emphasizes the completeness of inferred entities and their interrelations. To capture this aspect, we introduce evaluation metrics based on the coverage of entities and relations. Following the procedure used during training data construction, we extracted knowledge graphs from the natural language reasoning processes generated by the model, resulting in triplets in the form of Equation (1). We then compute the Jaccard similarity (Jaccard, 1912) between the entity and relation sets generated by the model during training and those present in the reference reasoning processes.

$$R_{statistic} = \max_{i\in[1,N]} \left(\underbrace{\frac{|\mathcal{E}_{\text{ref}}^{(i)} \cap \mathcal{E}_{\text{gen}}|}{|\mathcal{E}_{\text{ref}}^{(i)} \cup \mathcal{E}_{\text{gen}}|}}_{\text{Key Entity Coverage}} + \underbrace{\frac{|\mathcal{L}_{\text{ref}}^{(i)} \cap \mathcal{L}_{\text{gen}}|}{|\mathcal{L}_{\text{ref}}^{(i)} \cup \mathcal{L}_{\text{gen}}|}}_{\text{Relation Coverage}} \right) \quad (6)$$

- $R_{logical}$: Moreover, to evaluate the multi-step logical structure of reasoning process, we also assess the logical architecture alignment by computing the Jaccard similarity between the j-hop ($j = 1, 2, ..., K$) reasoning paths generated by the model during training, denoted as $\mathcal{P}_{\text{gen}}^{(j)}$, and the i-th reference diagnostic reasoning trajectories $\mathcal{P}_{\text{ref}}^{(ij)}$. Here K is the minimum path length between the longest reasoning chains extracted from the model-generated and reference medical knowledge graphs.

$$R_{logical} = \max_{i\in[1,N]} \frac{2}{K(K+1)} \sum_{j=1}^{K} j \cdot \frac{|\mathcal{P}_{\text{ref}}^{(ij)} \cap \mathcal{P}_{\text{gen}}^{(j)}|}{|\mathcal{P}_{\text{ref}}^{(ij)} \cup \mathcal{P}_{\text{gen}}^{(j)}|} \quad (7)$$

Retrieval Number. Another objective of this stage is to encourage the model to conduct external knowledge searches to assist its reasoning process, thereby laying the foundation for improving retrieval capabilities in the subsequent stage. To this end, we established a reward mechanism based on the number of search operations performed:

$$\mathcal{R}_{retrieval_num} = \begin{cases} 1, \text{if} & n \geq \delta \\ 0, \text{if} & n < \delta \end{cases} \quad (8)$$

where n indicates the number of retrieval calls and the minimum allowable search count is $\delta = 3$.

2.2.2 Stage 2: Awakening the Retriever

Following the initial training stage, the model has acquired a foundational paradigm for medical reasoning. In this stage, we shift our focus to enhancing the model's ability to retrieve external knowledge. Specifically, we aim to improve the generation of semantically accurate and retrieval-efficient query terms. In general, the design of the reward function in Stage 2 is *the sum of normalized* components: ***format***, ***answer correctness***, and ***retrieval effectiveness***, with the first two already formally defined in Equation (2) and Equation (3).

Retrieval Effectiveness. It encompasses two normalized reward components: ***(1)*** the authority of the retrieved medical documents, and ***(2)*** the extent to which the retrieved content is utilized throughout the entire reasoning process:

$$\mathcal{R}_{retrieval} = \tilde{R}_{quality} + \tilde{R}_{breadth} \quad (9)$$

- $R_{quality}$: Each retrieved document is assigned an integer evidence level e by the proprietary model (e.g., DeepSeek-V3) according to the principles of Evidence-Based Medicine (EBM) (Sackett et al., 1996), where $e \in \{x \in \mathbb{Z} \mid 1 \leq x \leq 6\}$, with 1 indicating the highest level of credibility. Here D denotes the number of documents retrieved by the model during a single rollout:

$$R_{quality} = \frac{1}{D} \sum_{j=1}^{D} (6 - (e_j - 1)) \quad (10)$$

- $R_{breadth}$: We also follow the procedure during data construction, where we extract knowledge graph triplets in the form of Equation (1) from the reasoning processes generated by the model. Therefore, the proportion of retrieved triplets in the reasoning process is:

$$R_{breadth} = \frac{1}{|\mathcal{G}|} \sum_{j=1}^{|\mathcal{G}|} s_j = \frac{\sum_{(h,l,t,s)\in\mathcal{G}} \mathbb{I}(s=1)}{|\mathcal{G}|} \quad (11)$$

2.2.3 Stage 3: Orchestrating the Dual-Process Collaboration

Having separately enhanced the model's reasoning and retrieval capabilities in the earlier stages, Stage 3 concentrates on strengthening the coordination between reasoning and retrieval. The reward function at this stage is *the sum of normalized* ***format*** and ***answer correctness***, which directly prioritizes end-to-end (E2E) performance of the ultimate objective, as defined in Equation (2) and Equation (3).

3 Experiments

3.1 Experimental Setup

Datasets. Datasets for *training* are detailed in the previous methodology section. For *evaluation*, we have selected seven medical datasets including the MedQA-USMLE, MedQA-MCMLE (Jin et al., 2020), MedMCQA (Pal et al., 2022), RareArena[2] (THUMedInfo, 2025), MMLU-Med (Hendrycks et al., 2021), NEJMQA (Katz et al., 2024) and MedXpertQA (Alonso et al., 2024), covering both standard and real-world clinical scenarios. We utilize LLM-as-Judge to verify the correctness of the responses, and then calculate the accuracy scores as the evaluation metric.

[2]For MedQA, we merge the dev and test subsets for evaluation. For MedMCQA, dev subset is employed for assessment. For RareArena, we use the test sets that we have split before.

Model	Method	MedQA-USMLE	MedQA-MCMLE	MedMCQA	RareArena-RDC	RareArena-RDS	MMLU-Med	NEJMQA	MedXpertQA	Avg.
		In-Domain (ID)					*Out-of-Domain (OOD)*			
Close-Sourced Models										
GPT-4o-mini	–	74.45	69.25	70.52	52.03	43.24	80.96	57.87	21.58	58.74
Open-Sourced Medical-Specific Models										
MEDITRON-7B	–	48.67	44.28	46.78	37.97	21.60	50.12	33.40	16.55	37.42
UltraMedical3-8B	–	61.57	52.42	61.82	40.76	28.54	72.52	45.31	10.82	46.72
UltraMedical3.1-8B	–	66.86	58.45	65.73	43.83	32.39	75.86	50.66	12.08	50.73
MEDITRON-70B	–	60.60	55.64	56.48	75.16	48.85	70.53	65.33	18.72	56.41
Open-Sourced Medical Reasoning Models										
HuatuoGPT-o1-8B	–	66.97	66.15	72.45	49.76	41.59	74.52	51.60	14.34	54.67
MedS^3-8B	–	73.51	69.63	65.47	46.85	37.72	78.75	55.09	12.50	54.94
AlphaMed-8B	–	64.06	64.98	66.43	43.55	38.14	71.44	51.48	22.01	52.76
Open-Sourced Base / Instruct Models										
LLaMA3.1-8B -Instruct	Naive Response	31.16	41.45	30.02	41.85	22.16	37.12	50.41	14.90	33.63
	SFT	61.39	62.10	63.27	44.83	35.20	63.08	49.64	12.72	49.03
	R1-Searcher∗	60.28	60.92	63.54	42.87	38.65	70.19	53.81	15.73	50.75
	ReSearch∗	62.76	66.03	66.25	46.35	38.44	71.27	53.26	14.65	52.38
	Med-R^3 (ours)	**75.91**	**75.95**	**75.89**	**57.34**	**47.16**	**79.07**	**60.60**	**16.48**	**61.05**
Qwen2.5-7B	Naive Response	22.58	39.14	28.77	32.17	23.10	44.45	41.48	11.59	30.41
	SFT	52.56	50.04	57.90	53.45	34.67	56.94	48.23	11.28	45.63
	R1-Searcher∗	56.78	49.70	58.35	53.69	33.27	66.81	52.98	12.55	48.02
	ReSearch∗	62.47	60.24	63.11	55.95	34.68	70.29	52.30	12.67	51.46
	Med-R^3 (ours)	**68.64**	**67.53**	**68.97**	**63.02**	**45.76**	**75.81**	**58.54**	**14.98**	**57.91**
Qwen2.5-14B	Naive Response	50.01	50.70	42.85	43.17	26.58	71.60	45.63	11.06	42.70
	SFT	68.85	70.22	70.15	75.27	52.82	75.81	47.08	11.54	58.97
	R1-Searcher∗	69.20	71.75	68.45	76.32	54.05	77.69	52.08	12.65	60.27
	ReSearch∗	69.52	74.05	72.20	75.67	53.54	80.25	50.65	13.10	61.12
	Med-R^3 (ours)	**78.01**	**80.59**	**75.42**	**77.94**	**58.15**	**85.33**	**62.40**	**15.69**	**66.69**

Table 1: Comparison of Med-R^3 with baselines. ∗ denotes the re-implementation with the same amount of our constructed data for a fair comparison. The best and second best of each model are in **bold** and underlined.

Medical Knowledge Corpus. We establish a comprehensive medical knowledge base to support information retrieval, curated from multiple data reliable sources (Roberts, 2001; Hoeppner, 2012; Chen et al., 2023; Foundation, 2024; Lu et al., 2025). It comprises four representative types of resources: *academic papers*, *entries*, *books*, and *guidelines*, offering both depth and breadth of information. We employ the hybrid retrieval strategy, including BGE-Large-EN-v1.5 (Xiao et al., 2023) for the dense and SPLADE-v3 (Lassance et al., 2024) for the sparse retrieval. During the rollout phase in both training and evaluation, we retrieve the top-5 related documents for each query. More details can be found in Section A.2.

Models and Implementation. We use LLaMA3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B and Qwen2.5-14B (Yang et al., 2024) as the backbone models for our training. The *reinforcement learning (RL)* framework is built on verl (Sheng et al., 2025) with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the learning algorithm. The training dataset obtained from the data construction phase contains 10,521 samples. Each sample undergoes 16 rollouts during training, with a training batch size of 256 and a rollout batch size of 64. The total number of training epochs is set to 3, where each training stage corresponds to 1 epoch comprising 41 training steps. The learning rate is 1e-6. Notably, in our training setup, external documents retrieved by the model are concatenated into the reasoning process, which are not generated by the training policy. To prevent these retrieved segments from influencing gradient updates, we apply masking during loss calculation, where we mask out all content enclosed within `<document>...</document>` tags. When conducting *supervised fine-tuning (SFT)* for method comparison, we utilized a learning rate scheduler featuring linear warm-up and cosine decay, with a learning rate peaking at 2e-5, alongside a warmup ratio of 0.03 and a weight decay of 0.0.

Baselines. We compare our Med-R^3 method against the following baselines: (1) We employ GPT-4o-mini (Hurst et al., 2024) as the ***Close-Sourced Models*** competitors. (2) ***Open-Sourced Medical-Specific Models*** include MEDITRON-7B, MEDITRON-70B (Chen et al., 2023), UltraMedical3-8B, and UltraMedical3.1-8B (Zhang et al., 2024). (3) ***Open-Sourced Medical Reasoning Models*** comprise HuatuoGPT-o1-8B (Chen et al., 2024b), MedS^3-8B (Jiang et al., 2025) and AlphaMed-8B (Liu et al., 2025b), the backbone models of which are all LLaMA3.1-8B-Instruct (Dubey et al., 2024). (4) The simplest baseline is ***Naive Response***, where the model generates

responses directly without external knowledge retrieval or dataset training. (5) We also compare with the training strategy in which we perform ***Supervised Fine-Tuning (SFT)*** using our constructed reasoning trajectories produced by frontier models. (6) ***General Retrieval-Augmented Reasoning RL*** for comparison includes R1-Searcher (Song et al., 2025) and ReSearch (Chen et al., 2025). Details can be found in Section C.1.

3.2 Main Results

The main results of Med-R^3 are demonstrated in Table 1, and we summarize the observations below.

Med-R^3 is effective across different models. Results in Table 1 show that Med-R^3 consistently outperforms other baseline methods on both base and instruction-tuned models across different scales in terms of medical problem-solving. Compared to *naive response*, Med-R^3 enhances average downstream task performances by 73.93%. Furthermore, models enhanced with Med-R^3 demonstrate the potential to outperform *close-sourced models* in medical scenarios. Specifically, LLaMA3.1-8B-Instruct + Med-R^3 achieves an average improvement of 3.93% over GPT-4o-mini at a comparable parameter scale, while Qwen2.5-14B + Med-R^3 shows a more substantial gain of 13.53%. Remarkably, even when compared to larger *open-sourced medical-specific models* such as MEDITRON-70B (Chen et al., 2023), our approach achieves over 9.70% higher performance on average with much smaller parameter scale (7B–14B).

Retrieval-augmented-reasoning boosts medical performances. The backbone model for all *open-sourced medical reasoning models* considered in this study is LLaMA3.1-8B-Instruct (Dubey et al., 2024). One of the advantages of Med-R^3 lies in its ability to dynamically retrieve and incorporate external medical knowledge during the reasoning process. Experimental results demonstrate that Med-R^3 achieves an average performance improvement of over 12.80% across medical tasks compared to these competitors. This finding further underscores the importance of integrating external knowledge retrieval during reasoning, particularly in the medical domain, where knowledge is highly specialized, rapidly evolving, and broad in scope.

4 Ablations and Analysis

Ablation studies and in-depth analysis are performed on models to highlight the necessity of progressive multi-stage training and the contribution of each training stage, as well as to assess the impact of their sequential order and the influence of medical-specific reward design on Med-R^3.

Order	In-Domain	Out-of-Domain	Avg.
Standard Pipeline			
1 → 2 → 3	**62.78**	**49.78**	**57.91**
Necessity of Progressive Training			
1 & 2 & 3	60.04 (↓ 4.36%)	47.00 (↓ 5.58%)	55.15 (↓ 4.77%)
The Role of Each Stage			
2 → 3	57.86 (↓ 7.84%)	46.71 (↓ 6.17%)	53.68 (↓ 7.30%)
1 → 3	59.28 (↓ 5.58%)	46.54 (↓ 6.51%)	54.50 (↓ 5.89%)
1 → 2	61.21 (↓ 2.50%)	47.95 (↓ 3.68%)	56.24 (↓ 2.88%)
Sequential Order of Stages			
2 → 1 → 3	<u>61.29</u> (↓ 2.37%)	<u>48.33</u> (↓ 2.91%)	<u>56.43</u> (↓ 2.56%)

Table 2: Analysis of training stages and sequential order using ***Qwen2.5-7B***. "Order": the sequence of each stage. The best and second best are in **bold** and <u>underlined</u>.

Method	In-Domain	Out-of-Domain	Avg.
Med-R^3	**62.78**	**49.78**	**57.91**
Analysis of $\mathcal{R}_{reasoning} = R_{semantic} + R_{statistic} + R_{logical}$			
w/o $R_{semantic}$	<u>61.78</u> (↓1.59%)	48.42 (↓2.73%)	<u>56.77</u> (↓1.97%)
w/o $R_{statistic}$	59.00 (↓6.02%)	47.37 (↓4.84%)	54.64 (↓5.65%)
w/o $R_{logical}$	60.95 (↓2.91%)	<u>48.67</u> (↓2.23%)	56.34 (↓2.71%)
Analysis of $\mathcal{R}_{retrieval} = R_{quality} + R_{breadth}$			
w/o $R_{quality}$	61.37 (↓2.25%)	47.74 (↓4.10%)	56.26 (↓2.85%)
w/o $R_{breadth}$	60.68 (↓3.35%)	47.60 (↓4.38%)	55.77 (↓3.70%)

Table 3: Analysis of the reward design using ***Qwen2.5-7B***. The best and second best are in **bold** and <u>underlined</u>.

4.1 Necessity of Progressive Training

We consolidated the reward functions from all stages throughout training to verify the importance of progressively optimizing the model's retrieval and reasoning capabilities in a staged manner. The results are presented in Table 2 (1 & 2 & 3), where we observe a performance drop of 4.77% compared to our original multi-stage training strategy (1 → 2 → 3). A key reason for this decline is the inherent complexity and potential conflicts among diverse reward signals, i.e., the retrieval-focused and reasoning-oriented components impose different behavioral pressures on the model, leading to unstable policy updates in training. For instance, early in training, the model may not yet possess a sufficiently developed reasoning structure to effectively utilize or prioritize external knowledge retrieval, resulting in misaligned gradient signals and diminished learning efficiency.

4.2 The Role of Each Stage

To evaluate the significance of each stage, we conduct three ablation experiments by individually removing Stage 1, 2, and 3. We then assess the performance on benchmarks, including standard as well as real-world clinical scenarios, with results summarized in Table 2. To control for the influence of training data volume on model performance, we set the total number of training epochs at 3 to align with the main experiment, and allocate 1.5 epochs to each of the remaining two stages for training. Detailed analysis is provided in Section C.4.1.

4.3 Sequential Order of Stages

We swap Stage 1 and Stage 2 to assess their impact on models. As shown in Table 2 ($2 \rightarrow 1 \rightarrow 3$), exchanging Stage 1 and Stage 2 leads to a slight performance drop by 2.56%. Our analysis of each stage demonstrates the robustness and rationale of our original training sequence, which first optimizes reasoning capabilities, followed by adaptive refinement of retrieval abilities. It stems from the fact that a solid reasoning foundation is essential to recognize retrieval needs during problem-solving. Only with this prerequisite can the model meaningfully improve its ability to generate retrieval queries that effectively capture relevant information.

4.4 Analysis of Reward Design

We analyzed the components of the two reward metrics, $\mathcal{R}_{reasoning}$ and $\mathcal{R}_{retrieval}$, with results summarized in Table 3. The findings indicate that the absence of $R_{statistic}$ in $\mathcal{R}_{reasoning}$ leads to a higher drop in model performance. This highlights that the comprehensive coverage of medical entities and relations plays a crucial role in the reasoning process, which is a key distinction from reasoning processes in mathematics and code. Furthermore, the removal of $R_{breadth}$ from $\mathcal{R}_{retrieval}$ also results in a notable decline in performance, underscoring the importance of evaluating the contribution of the retrieved documents to the reasoning process.

5 Related Work

LLMs for Medical Domain Large Language Models (LLMs) have been increasingly deployed in the medical field (Zeng et al., 2020; Gu et al., 2021; Clusmann et al., 2023). Extensive studies have focused on the direct use of medical data for the pretraining or supervised fine-tuning of LLMs (Singhal et al., 2023; Thirunavukarasu et al., 2023), leading to prominent open-source milestones such as MEDITRON (Chen et al., 2023) and UltraMedical (Zhang et al., 2024) series. However, solving medical problems requires structured multi-step reasoning (Lucas et al., 2024; Savage et al., 2024). Existing research indicates that the pretraining and SFT phases diminish the model's generalization when confronted with novel scenarios (Havrilla et al., 2024; Lai et al., 2025). In contrast, the reinforcement learning (RL) phase is more conducive to cultivating a model's cognitive abilities, particularly in domains that demand substantial logical reasoning (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025). Nevertheless, unlike math and coding skills, the specific knowledge inherent in the medical domain is not always available within foundation models (Singhal et al., 2023; Wang et al., 2023). Therefore, establishing connections to the knowledge corpus to acquire external information is of paramount importance for medical scenarios.

Reinforcement Learning in LLMs Compared to supervised fine-tuning (SFT), reinforcement learning (RL) provides an alternative by enabling emergent reasoning of models without explicit supervision (Jaech et al., 2024; Team et al., 2025). Efforts have been made to improve the medical reasoning capability in LLMs via the RL process, with notable works including HuatuoGPT-o1 (Chen et al., 2024b), Med-S^3 (Jiang et al., 2025), and AlphaMed (Liu et al., 2025b). However, they neglect the fact that specialized medical knowledge is not sufficiently encoded within the model's learned parameters. Therefore, the integration of external knowledge becomes crucial (Wang et al., 2024). While prior works focused on enhancing retrieval-augmented reasoning via RL in general-domain settings (Chen et al., 2025; Song et al., 2025), the reward modeling strategies developed therein are not well adapted to the clinical scenario. To mitigate this, we systematically employ the RL process to jointly enhance models' retrieval and reasoning capabilities in the medical domain.

6 Conclusion

In this paper, we introduce Med-R^3, a novel progressive reinforcement learning framework aimed at enhancing the medical retrieval-augmented reasoning capabilities of models. Extensive experimental results indicate that Med-R^3 achieves state-of-the-art training outcomes among open-sourced base and instruct models in medical scenarios.

7 Limitations

There are some limitations in our work. The construction of reference reasoning trajectories in the training data, as well as the evaluation of model performance during training (e.g., triplet extraction, evidence-level labeling, and reward calculation), relies on advanced large language models. This introduces a dependency and may lead to the propagation of biases.

8 Ethical Considerations

While deploying and utilizing Med-R^3 framework holds potential in medical contexts, it also brings several ethical considerations that must be addressed to ensure responsible and beneficial use. The external knowledge corpora and medical data involved in retrieval ensure data anonymization, encryption, and restricted access to authorized personnel throughout the retrieval and reasoning workflows. Additionally, Med-R^3 adhere to healthcare regulations and ethical guidelines. Its development and deployment are overseen by medical and ethical review bodies to ensure alignment with industry standards and patient safety imperatives.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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*.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, and 1 others. 2025. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

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

Wikimedia Foundation. 2024. Wikimedia downloads.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *Jmir Ai*, 4:e58670.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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*.

Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. 2016. Syntactic analyses and named entity recognition for PubMed and PubMed central — up-to-the-minute. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 102–107, Berlin, Germany. Association for Computational Linguistics.

Peter Hall. 1987. On kullback-leibler loss and density estimation. *The Annals of Statistics*, pages 1491–1519.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Marilu A Hoeppner. 2012. Ncbi bookshelf: books and documents in life sciences and health care. *Nucleic Acids Research*, 41(D1):D1251–D1260.

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*.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

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*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds^3: Towards medical small language models with self-evolved slow thinking. *arXiv preprint arXiv:2501.12051*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Uriel Katz, Eran Cohen, Eliya Shachar, Jonathan Somer, Adam Fink, Eli Morse, Beki Shreiber, and Ido Wolf. 2024. Gpt versus resident physicians—a benchmark based on official board scores. *Nejm Ai*, 1(5):AIdbp2300192.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. Splade-v3: New baselines for splade. *arXiv preprint arXiv:2403.06789*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *NeurIPS*, 35:3843–3857.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025a. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025b. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.

Keer Lu, Zheng Liang, Zhuoran Zhang, Da Pan, Shusen Zhang, Xin Wu, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, Tengjiao Wang, and Wentao Zhang. 2025. Med-r^2: Crafting trustworthy llm physicians through retrieval and reasoning of evidence-based medicine. *arXiv preprint arXiv:2501.11885*.

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.

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.

Richard J Roberts. 2001. Pubmed central: The genbank of the published literature.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, and 1 others. 2021. Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10–D17.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

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*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

THUMedInfo. 2025. Rarearena: A dataset for rare disease information retrieval.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pretrained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Wikimedia Foundation. 2024. Wikipedia dataset. https://huggingface.co/datasets/wikimedia/wikipedia. Accessed: 2025-07-08.

Xuehong Wu, Junwen Duan, Yi Pan, and Min Li. 2023. Medical knowledge graph: Data sources, construction, reasoning, and applications. *Big Data Mining and Analytics*, 6(2):201–217.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

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*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, and 1 others. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Appendix

A Data Preprocessing Details **12**
- A.1 Training Datasets Details 12
- A.2 Knowledge Corpus Details 12

B Progressive Reinforcement Learning Details **13**
- B.1 Preliminaries 13
- B.2 Reward Modeling Details 15

C Experiment Details **17**
- C.1 Baselines 17
- C.2 Benchmarks 18
- C.3 More Experimental Observations . 18
- C.4 Ablation Study and Analysis Details 19
- C.5 Environment and Hardware Configurations 19

D Case Studies **19**

E Prompts **23**

A Data Preprocessing Details

A.1 Training Datasets Details

We have collected the following datasets for our training, covering both closed-set exam questions and rare disease diagnostic scenarios.

- **MedQA-USMLE** and **MedQA-MCMLE** (Jin et al., 2020): The MedQA dataset is sourced from the professional medical board exams, including content in English, simplified Chinese, and traditional Chinese. In this study, we use the subsets from the US and Mainland China. MedQA-USMLE is based on 18 English medical textbooks, while MedQA-MCMLE is constructed from 33 simplified Chinese textbooks.

- **MedMCQA** (Pal et al., 2022): This dataset is a large-scale collection of multiple-choice questions covering 2,400 health-related topics and 21 medical disciplines. The content spans medicine (such as endocrinology, infectious diseases, haematology, and respiratory medicine), surgery (including general surgery, endocrinological surgery, breast and vascular surgery), as well as radiology and biochemistry. The questions are derived from both real-world clinical scenarios and simulated examinations.

- **RareArena** (THUMedInfo, 2025): This dataset is a comprehensive dataset curated for rare disease diagnosis, encompassing nearly 50,000 patient cases that cover more than 4,000 diseases. It features 2 task settings, *Rare Disease Screening (RDS)* where patient records are truncated before any diagnostic tests are performed, and *Rare Disease Confirmation (RDC)* where cases are truncated at the point of final diagnosis.

The data construction pipeline can be found in Figure 3. After data filtering, we constructed a final training dataset consisting of 2,140 questions from MedQA-USMLE, 1,204 from MedQA-MCMLE, 6,748 from MedMCQA, and 429 from RareArena. The training process was structured into 3 stages, with each stage trained for one full epoch.

For the *MedQA-USMLE* and *MedQA-MCMLE* datasets, the original data is divided into three parts: train, dev, and test. We utilize the training part for RL, and merge the dev and test subsets for evaluation. Since the test split of *MedMCQA* does not include ground truth labels, we use the training set for RL and reserve the development set for evaluation purposes. As for *RareArena*, we manually partitioned the RareArena-RDC task set into training and test sets at an 8:2 ratio using random sampling. It is worth noting that due to the distribution of question complexity, the majority of data filtered out from MedQA and MedMCQA were classified as overly simplistic, whereas in the case of RareArena, most excluded samples were considered overly complex.

A.2 Knowledge Corpus Details

We have established a comprehensive medical knowledge base to support document retrieval during training and evaluation, which comprises four representative types of resources: *academic papers*, *entries*, *books*, and *guidelines*.

- **Academic Papers** Academic literature provide valuable insights drawn from recent scientific research, offering a strong theoretical basis for guiding clinical practice and public health policies. We sourced the publicly available academic articles from PubMed Central (PMC) (Roberts, 2001), and processed them following the pipeline of Hakala et al. (2016).

- **Entries** Medical entries serve as a rich source of multidimensional healthcare information, spanning clinical applications to biomedical

Source Type	Data Resource	#Volume	#Avg. Token
Papers	PubMed Central (Roberts, 2001)	600,000	3,820
Entries	Wikipedia (Foundation, 2024)	470,000	1,387
Books	NCBI Bookshelf (Hoeppner, 2012)	10,000	4,083
Guidelines	Guidelines from Chen et al. (2023) when training MEDITRON series	10,000	1,100

Table 4: Overall statistics of medical knowledge resources.

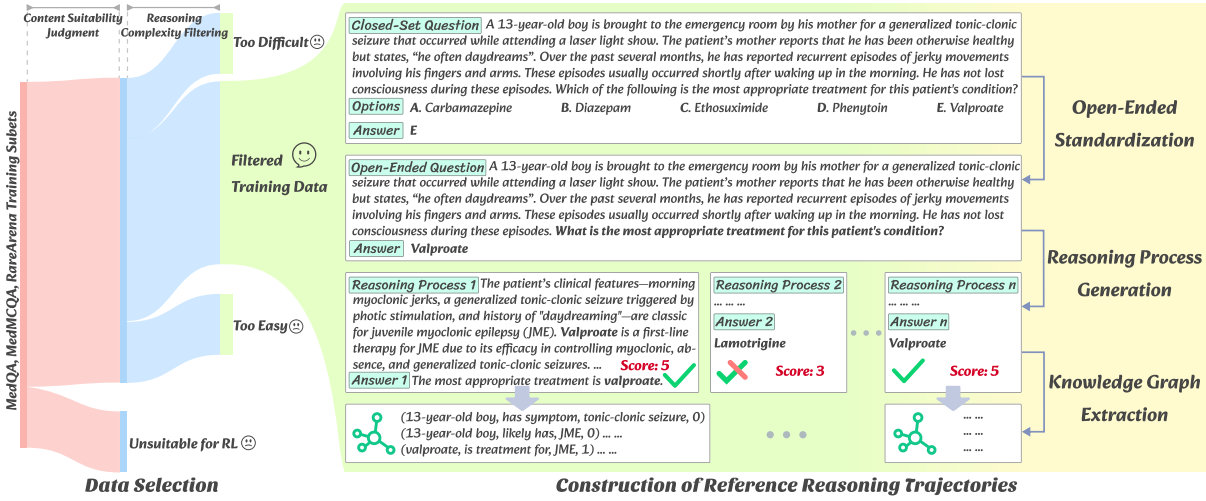

Figure 3: The pipeline of training data construction for Med-R^3.

research. We compile such entries by extracting and curating health-related content from the Wikipedia dataset (Wikimedia Foundation, 2024) to build the entries. The final collection comprises approximately 470k documents.

- **Books** Medical textbooks are key sources of structured and validated medical knowledge, valuable for tackling complex clinical problems or staying updated on specific diseases. We compiled materials from the NCBI Bookshelf (Sayers et al., 2021), and following the handling strategy of Hakala et al. (2016), finally obtained 10k documents for knowledge retrieval.

- **Guidelines** Clinical practice guidelines play a critical role in Evidence-Based Medicine (EBM) by providing evidence-informed recommendations to guide clinical decision-making. We incorporated the guideline data when training the MEDITRON series (Chen et al., 2023), and curated approximately 10k documents.

The statistical information on volume and token number across knowledge resources can be found in Table 4. For retrieval, we segment the texts into chunks containing no more than 1,000 tokens. The segmentation prioritizes natural divisions such as chapters or sections. When such structural boundaries are unavailable or exceed the token limit, we apply truncation based on the predefined threshold to ensure consistency in input length.

B Progressive Reinforcement Learning Details

Here, we first introduce the core concepts and methodologies employed in our progressive reinforcement learning process, and then provide an elaboration on the algorithms and procedural significance across the different stages.

B.1 Preliminaries

B.1.1 Group Relative Policy Optimization (GRPO)

During the training process, we utilize the Group Relative Policy Optimization (GRPO) as the RL algorithm. For each question $x \sim \mathcal{D}$, the behavior policy $\pi_{\theta_{\text{old}}}$ generates a set of G candidate completions $\tau = \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$, with each response receiving a scalar reward r_i. The training objective is to optimize the policy π_θ based on reference policy $\pi_{\theta_{\text{ref}}}$:

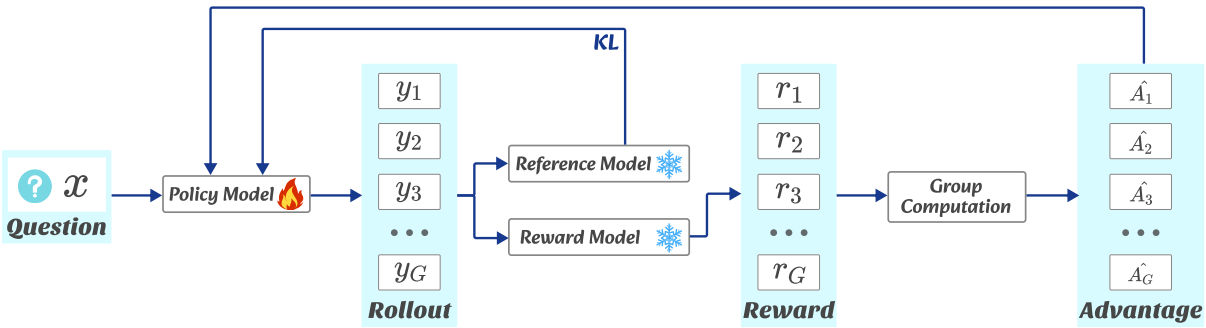

Figure 4: The Group Relative Policy Optimization (GRPO) pipeline.

$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \frac{1}{G}\sum_{i=1}^{G}[\min(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}\hat{A}_i, \text{clip}(\frac{\pi_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)}, 1-\epsilon, 1+\epsilon)\hat{A}_i) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\theta_{\text{ref}}})] \tag{12}$$

where the group-normalized advantage $\hat{A}_i$ of the i-th rollout in current group is defined as:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}$$

The illustration of GRPO is shown in Figure 4. Here ϵ is the clipping ratio, a hyperparameter controlling the tolerance for policy deviation, and the `clip` function clips the importance weight r_i to the interval $[1-\epsilon, 1+\epsilon]$, which stabilizes training and mitigating the risk of policy collapse. β is the Kullback–Leibler (KL) loss coefficient (Hall, 1987). To ensure stability during policy updates, a KL divergence penalty is included in the optimization objective, preventing significant deviations from the original reference policy models.

B.1.2 Jaccard Similarity

Jaccard similarity (Jaccard, 1912) is a widely used measure of similarity between two sets, quantifying the degree of overlap by comparing shared and distinct elements. It is defined as the ratio of the size of the intersection to the size of the union of the two sets. Formally, given two sets A and B, the Jaccard similarity is expressed as:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{13}$$

where $|A \cap B|$ denotes the number of elements common to both sets, and $|A \cup B|$ represents the total number of distinct elements in either set. The Jaccard similarity ranges from 0, indicating no overlap, to 1, indicating identical sets. It is applicable

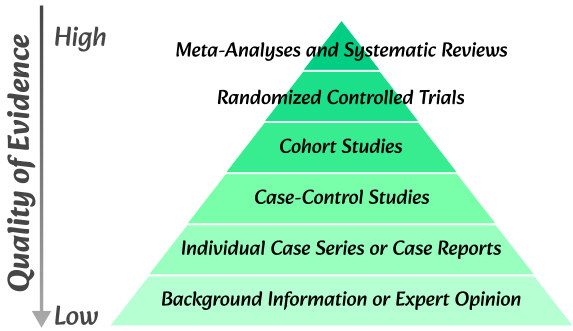

Figure 5: Illustration of *Hierarchy of Evidence* in EBM.

to sets containing either numerical values or categorical strings. Here we use Jaccard similarity to compute the reward signals during Stage 1 of deliberative reasoner cultivation, specifically for quantifying the overlap in reasoning paths used to derive $R_{statistic}$ and $R_{logical}$.

B.1.3 Evidence-Based Medicine (EBM)

It involves the integration of individual expertise with the most reliable external clinical evidence derived from systematic research (Sackett et al., 1996). To identify the "best evidence", researchers evaluate trial quality using grading systems that assess the likelihood of bias and the reliability of results. The ***hierarchy of evidence*** informs clinical decision-making by prioritizing evidence according to its methodological strength, as illustrated in Figure 5. Its structure is typically organized from the highest to the lowest level of reliability, informing the quality scoring reward ($R_{quality}$) for the retrieved documents in Equation (10).

- **Systematic Reviews/Meta-Analyses (SR/MA)** represent *the highest* level of evidence, evaluating the consistency and risk of bias across medical studies to summarize the overall effect of interventions or exposures.

- **Randomized Controlled Trials (RCTs)** constitute *the second-highest* level of evidence, de-

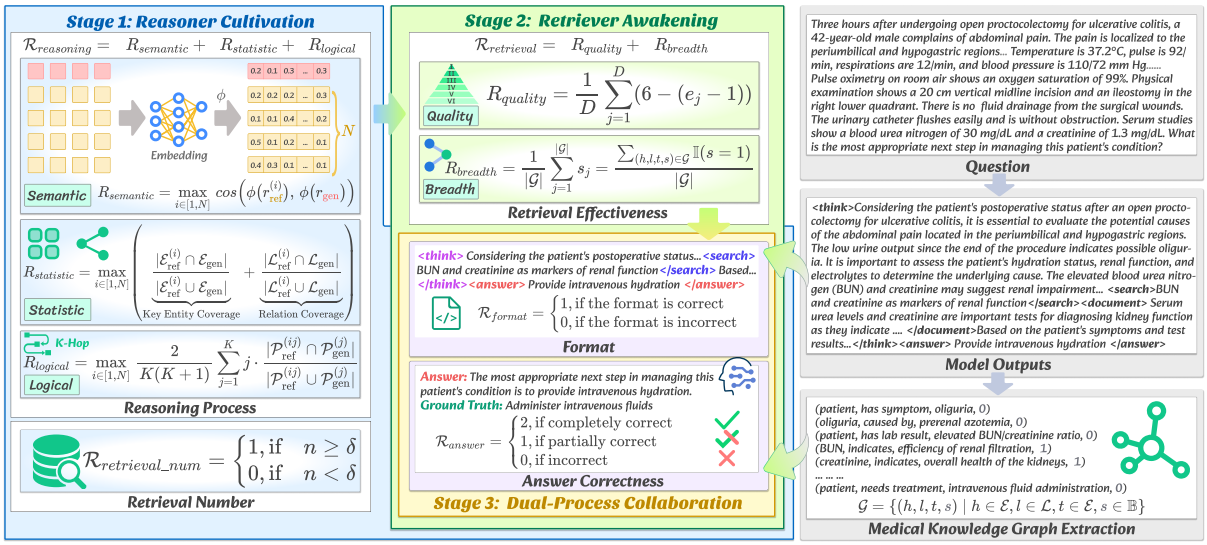

Figure 6: Detailed illustration of the reward function design during the progressive reinforcement learning of Med-R^3. The model's natural language outputs are first structured into knowledge graph representations. In each training stage, tailored reward functions assess both the textual answer and the corresponding reasoning trajectory in the format of Equation (1).

signed to minimize confounding biases and assess causal relationships between interventions and outcomes across study groups.

- **Cohort Studies** represent *the third-highest* level of evidence. While both retrospective and prospective designs are subject to bias, prospective studies are generally more reliable, offering better control over information bias.
- **Case-Control Studies** rank as *the fourth-highest* level of evidence. They aim to identify associations between outcomes and past exposures but are prone to selection, information, and confounding biases, which limit their reliability compared to cohort studies.
- **Individual Case Series and Case Reports** are of *the second-lowest* level, essentially consisting of uncontrolled studies without a comparison group. The lack of controls limits the ability to establish reliable associations between interventions, exposures, or risk factors and outcomes.
- **Expert Opinion** is considered *the lowest* level of evidence due to its inherent susceptibility to bias. Experts may favor information that supports existing beliefs, leading to confirmation bias, conflicts of interest, and a narrow focus that overlooks broader contextual factors.

B.2 Reward Modeling Details

We provide an in-depth description of the reward computation process associated with progressive RL described in the main content. As shown in the right part of Figure 6, during the verification phase of each training stage, we first parse the model's natural language output into a structured knowledge graph representation in the form of Equation (1). The stage-specific reward metric is then applied to assess both the *natural language response* and the structural and logical accuracy of the *reasoning trajectory encoded in the knowledge graph*.

B.2.1 Stage 1: Reasoner Cultivation

The primary objective of this stage is twofold: ***(1)*** to regularize the model's output format, ensuring structured and consistent responses, and ***(2)*** to cultivate its ability to perform logical reasoning when answering medical questions. In addition, by incorporating reward signals based on the number of knowledge retrievals, we encourage the model to make increased use of external medical knowledge sources, which lays the groundwork for enhancing the model's retrieval capabilities in the subsequent stage. The reward is composed of four aspects: *format*, *answer correctness*, *reasoning process* and *retrieval number*, as depicted in Figure 6. While format and answer correctness are commonly used reward components, we place particular emphasis on the rationale behind the design of rewards for reasoning process and retrieval number.

Reasoning Process. In designing the reward function, we have considered the following aspects:

- **Alignment with High-Quality Reasoning Trajectories**: We aim to guide the model's reasoning

Algorithm 1 Evaluation for the Logical Structure of the Reasoning Process (details for Equation (7))

Input: Model-generated reasoning paths $\mathcal{P}_{\text{gen}}$, reference reasoning trajectories $\mathcal{P}_{\text{ref}}$
Parameter: Maximum path length K for comparison
Output: Logical score of the reasoning process R_{logical}

Initialize logical score: $R_{\text{logical}} \leftarrow 0$
Compute maximize path length:

$$K \leftarrow \min\{\max_{p \in \mathcal{P}_{\text{gen}}} |p|, \max_{p \in \mathcal{P}_{\text{ref}}} |p|\}$$

for each reference trajectory $\mathcal{P}_{\text{ref}}^{(i)} \in \mathcal{P}_{\text{ref}}$ **do**
  Initialize the logical score for $\mathcal{P}_{\text{ref}}^{(i)}$: $\text{Score}^{(i)} \leftarrow 0$
  for $j = 1, 2, ..., K$ **do**
    Filter model paths of length j:

$$\mathcal{P}_{\text{gen}}^{(j)} \leftarrow \{p \in \mathcal{P}_{\text{gen}} \mid |p| = j\}$$

$$\mathcal{P}_{\text{ref}}^{(ij)} \leftarrow \{p \in \mathcal{P}_{\text{ref}}^{(i)} \mid |p| = j\}$$

    if $\mathcal{P}_{\text{gen}}^{(j)} \neq \emptyset$ **then**
      Compute Jaccard similarity for the j-hop paths and update weighted score:

$$\text{Score}^{(i)} \leftarrow \text{Score}^{(i)} + j \cdot \frac{|\mathcal{P}_{\text{ref}}^{(ij)} \cap \mathcal{P}_{\text{gen}}^{(j)}|}{|\mathcal{P}_{\text{ref}}^{(ij)} \cup \mathcal{P}_{\text{gen}}^{(j)}|}$$

    end if
  end for
  Normalize the logical score for $\mathcal{P}_{\text{ref}}^{(i)}$:

$$\text{Score}^{(i)} = \text{Score}^{(i)} \cdot \frac{1}{\sum_{j=1}^{K} j} = \text{Score}^{(i)} \cdot \frac{2}{K(K+1)}$$

  Update $R_{\text{logical}} \leftarrow \max\{R_{\text{logical}}, \text{Score}^{(i)}\}$
end for
return R_{logical}

process toward high-quality trajectories observed in frontier models. To this end, our evaluation criteria incorporate semantic and structural characteristics of reference reasoning paths (see Equation (4) to Equation (7)), enabling the model to learn from expert-like inference patterns.

- **Capturing Medical Reasoning Specifics**: Medical reasoning often requires both logical coherence and comprehensive coverage of relevant entities and their relationships. Unlike mathematical or programming tasks that emphasize rigid logical order, medical reasoning allows for multiple valid pathways, provided they are semantically rich and clinically grounded (Wu et al., 2023; Gao et al., 2025). Accordingly, we introduce: ***(1)*** a structural coherence score based on the sequential logic of the reasoning path, as outlined in Equation (7), and ***(2)*** a completeness score evaluating the coverage of key entities and relations, as shown in Equation (6).

Model	Method	Average Retrieval Frequency		
		Stage 1	Stage 2	Stage 3
Qwen2.5-7B	w/o $\mathcal{R}_{retrieval_num}$	0.44	0.83	0.92
	w/ $\mathcal{R}_{retrieval_num}$	3.67	4.81	4.56

Table 5: Comparison of average retrieval frequency with and without $\mathcal{R}_{retrieval_num}$ in Stage 1.

- **Robust Evaluation via Maximum Similarity Matching**: For each training instance, we provide multiple reference reasoning trajectories. However, these references may not be highly similar to one another due to the diversity of valid clinical reasoning. Therefore, in designing Equation (5) to Equation (7), we opt to use the maximum similarity score between the model's reasoning process and the reference reasoning processes, both in terms of semantics and structure, rather than an average score, to ensure robust and meaningful scoring.

- **Flexibility and Generalization**: Leveraging the strengths of reinforcement learning (RL), we avoid rigidly constraining the model's reasoning trajectory. Instead, we encourage it to approximate high-quality reasoning behaviors while allowing room for variation and creativity. This flexible guidance enhances the model's generalization capability when faced with novel or out-of-distribution reasoning tasks. To support this goal, we design a multi-dimensional reward scheme that captures various aspects of medical logic reasoning, as depicted in Equation (4).

As detailed in Algorithm 1, we have also provided the algorithm implementation of $R_{logical}$ defined in Equation (7).

Retrieval Number. One of the key motivations for incorporating this reward in Stage 1 is to lay the foundation for enhancing the model's retrieval capabilities in the subsequent stage. Specifically, only when the model develops the habit of querying external knowledge upon encountering unfamiliar concepts during reasoning, can there be meaningful room for optimizing its retrieval behavior. Our experiments also confirm this intuition: without such a reward signal in Stage 1, certain models tend to rely solely on internal knowledge and cease performing external retrievals altogether. As observed in Table 5, using Qwen2.5-7B as an ex-

ample, the average number of knowledge retrievals during logical reasoning drops significantly when the retrieval count component is omitted from the reward function.

B.2.2 Stage 2: Retriever Awakening

The principal objective of this stage is to further enhance the model's ability to retrieve external knowledge built upon the medical reasoning capabilities developed in Stage 1. Specifically, we aim to improve the model's capacity to generate search queries such that the resulting documents ***(1)*** exhibit higher semantic relevance, and ***(2)*** contribute more meaningfully to the logical reasoning process. In other words, this stage focuses on adaptively optimizing the content within the `<search>`...`</search>` tags.

The core reward metric in this phase is the *retrieval effectiveness*, which evaluates both the quality of the retrieved documents and their influence within the model's overall reasoning trajectory, as outlined in Equation (9). For document quality assessment, we adopt the Evidence-Based Medicine (EBM) grading framework illustrated in Figure 5 to classify the retrieved documents into different quality levels. To evaluate the contribution of these retrieved documents to the model's reasoning trajectory, we compute the ratio of retrieved knowledge graph triples relative to the total number of triples in the reasoning trajectory, as defined in Equation (11). This metric reflects the extent to which external knowledge contributes to the logical inference process.

B.2.3 Stage 3: Dual-Process Collaboration

After the previous two stages which have separately enhanced the model's reasoning and retrieval cabilities, in the final training stage, the model's end-to-end (E2E) performance on medical problem-solving is directly optimized based on the accuracy of the final answers. As a result, the reward function in this stage consists solely of *format* and *answer correctness*.

C Experiment Details

C.1 Baselines

In this section, we provide a comprehensive overview of the various models as well as methods that serve as baselines in our comparative analysis.

- **Close-Sourced Models**: Close-sourced models are regarded as embodying the current peak performance across various capabilities of LLMs, and serve as the strongest baselines. Here we have selected GPT-4o-mini (Hurst et al., 2024) for comparison.
- **Open-Sourced Medical-Specific Models**: These models refer to domain-specific models that were trained specifically on medical data. We have selected MEDITRON-7B, MEDITRON-70B (Chen et al., 2023), UltraMedical3-8B, and UltraMedical3.1-8B (Zhang et al., 2024) to represent the open-sourced medical-specific models for comparison to assess Med-R^3's relative advantage. Specifically, the backbone model of UltraMedical3.1-8B is LLaMA3.1-8B-Instruct, which facilitates a more intuitive comparison.
- **Open-Sourced Medical Reasoning Models**: Inspired by the breakthrough of OpenAI o1, these models are capable of complex medical reasoning, applying o1-like methods to the medical field. We have selected HuatuoGPT-o1-8B (Chen et al., 2024b), MedS^3-8B (Jiang et al., 2025) and AlphaMed-8B (Liu et al., 2025b) as competitors. HuatuoGPT-o1-8B improved models capabilities in medical reasoning by performing a two-stage SFT + RL training on verifiable medical problems. MedS^3-8B equips the model with a self-evolution paradigm and proposes PRM-guided Vote-Sum (P-VS) strategy during inference to enhance long-chain reasoning capabilities in the medical domain. AlphaMed-8B is trained exclusively on rule-based RL. The backbone models of these competitors are all LLaMA-3.1-8B-Instruct, allowing for a more direct comparison with LLaMA-3.1-8B-Instruct + Med-R^3.
- **Naive Response**: It refers to the case where the model directly generates answers to the medical questions without training or retrieval from external knowledge bases.
- **Supervised Fine-Tuning (SFT)**: The model undergoes SFT using our constructed training dataset, with reasoning trajectories produced by frontier models (e.g., DeepSeek-V3), enriched with knowledge obtained through retrieval mechanisms. During training, retrieved documents are with masked losses.
- **General Retrieval-Augmented Reasoning RL**: We also compare our method with approaches that enhance the model's retrieval-augmented reasoning capabilities through RL in the general

domain. R1-Searcher (Song et al., 2025) is a two-stage outcome-based RL method designed to enhance the model's capabilities of searching and integrating additional knowledge during the reasoning process. To ensure a more equitable comparison, we replaced the training dataset and the knowledge corpus for retrieval with one that matches our experimental setup. In the first stage, we trained the model for 1 epoch, and in the second stage for 2 epochs, with a total number of training samples matching that of Med-R^3. ReSearch (Chen et al., 2025) considers search operations as integral components of the reasoning chain, and trains models to reason with search via RL without using any supervised data on reasoning steps. Here we also train the model for 3 epochs to align with our experimental configuration. The RL process is implemented by the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) algorithm.

We evaluate these baselines under different inference methods, including ***Naive*** (direct generation), ***CoT*** (pure reasoning), ***RAG*** (pure retrieval), and ***CoT-RAG*** (interleaved reasoning and retrieval), as detailed in Table 6.

C.2 Benchmarks

To evaluate the performance of our proposed Med-R^3 in both standard and real-world clinical scenarios of the medical domain, we have selected seven medical datasets, including the MedQA-USMLE, MedQA-MCMLE (Jin et al., 2020), MedMCQA (Pal et al., 2022), MMLU-Med (Hendrycks et al., 2021), RareArena (THUMedInfo, 2025), MedXpertQA (Alonso et al., 2024) and NEJMQA (Katz et al., 2024). Among these benchmarks, MedQA series, MedMCQA and MMLU-Med are *standard question-answering tasks*, while RareArena, MedXpertQA and NEJMQA are related to *real-world clinical scenarios*. We employ LLM-as-Judge based on the frontier model DeepSeek-V3 (Liu et al., 2024) to assess the correctness of the responses, then compute the accuracy scores as our evaluation metric.

We have introduced within-domain tasks in the section of *training datasets details*, where we use the training subsets for the RL process and test subsets (if no ground-truth labels provided, we use the development subsets) for in-domain evaluation. In the following, we give descriptions of the remaining out-of domain benchmarks:

- **MMLU-Med** (Hendrycks et al., 2021): The MMLU-Med task is composed of six medical domains, anatomy, clinical knowledge, professional medicine, human genetics, college medicine, and college biology, which is extracted from the MMLU benchmark.
- **NEJMQA** (Katz et al., 2024): The NEJMQA dataset is constructed from clinical case challenges sourced from The Lancet and the New England Journal of Medicine, which is centered on diagnostic reasoning using patient symptom information.
- **MedXpertQA** (Alonso et al., 2024): The MedXpertQA dataset incorporates specialty-specific evaluations and realistic clinical case questions derived from authentic medical practice. Here we utilize the *text part* of this benchmark.

C.3 More Experimental Observations

The performance of each baseline reported in Table 1 in the main content represents the best result achieved in four different inference strategies. The comprehensive scores across all inference strategies are provided in Table 6. In addition, we use the radar chart to visualize the performance of the methods that share the same backbone model of ***LLaMA3.1-8B-Instruct***, as shown in Figure 7. Furthermore, we summarize addition observations and insights derived from our experiments as follows:

Retrieval and reasoning capabilities need to be cultivated. When employing the *CoT-RAG* inference method, the performance improvements across different models relative to the *naive responses* vary significantly. Smaller-scale models such as Qwen2.5-7B and LLaMA3.1-8B-Instruct exhibit only modest gains under the *CoT-RAG* setup, with average improvements of just 4.11% and 6.81%, respectively, in the absence of targeted training. In contrast, the larger Qwen2.5-14B model demonstrates a stronger adaptation to *CoT-RAG*, achieving a more substantial improvement of 18.43%. This indicates that lightweight models, stand to benefit greatly from dedicated training aimed at enhancing their retrieval-augmented reasoning capabilities in the medical domain.

RL boosts generalizable medical performances. We compare the performance of models after SFT and RL phase. Specifically, R1-Searcher, ReSearch, and Med-R^3 all employ RL-based training approaches. As shown in Table 6, on the in-domain benchmarks, the performance gap between

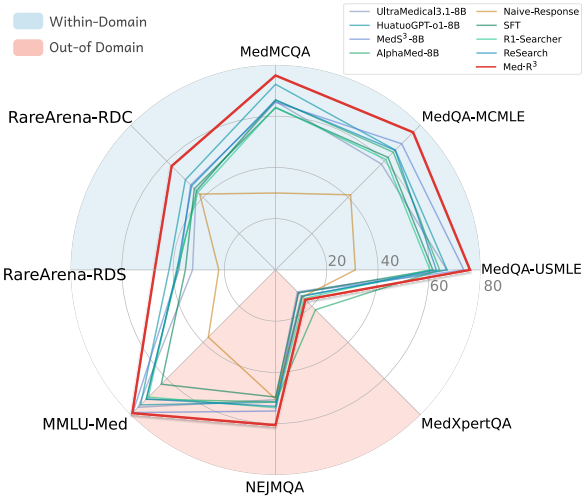

Figure 7: Comparison of Med-R^3 with baselines that utilize the ***LLaMA3.1-8B-Instruct*** as the backbone model.

SFT and R1-Searcher/ReSearch is relatively small, with SFT underperforming by 0.59% and 5.39%, respectively. However, on out-of-domain tasks, particularly on MMLU-Med, the gap becomes much larger, with SFT lagging behind R1-Searcher and ReSearch by 9.21% and 10.07%, respectively. A primary reason for this discrepancy lies in the nature of SFT as a strongly supervised learning paradigm. Under this setting, the model tends to learn by directly imitating the given reasoning trajectories, effectively memorizing problem-solving patterns rather than truly learning the underlying reasoning process. Consequently, its generalization capability is notably limited when faced with novel or unseen scenarios. It is worth noting that in our SFT setup, the training data includes retrieval-augmented reasoning trajectories that incorporate externally retrieved knowledge, which mitigates the inherent limitations of SFT to some extent. Nevertheless, a clear performance gap still remains.

Tailored reward design counts for medical scenarios. Compared to general outcome-based approaches that enhance retrieval-augmented reasoning through reinforcement learning (RL), e.g., R1-Searcher and ReSearch, our Med-R^3 achieves an average improvement of 16.73% and 12.54% in medical problem-solving. This performance gain is primarily attributed to our effective supervision of the model's medical reasoning process and the design of reward functions specifically tailored to the medical domain.

C.4 Ablation Study and Analysis Details

We report the original performance scores of the models on each benchmark during training stage ablation as well as reward design ablation, as depicted in Table 7 and Table 8.

C.4.1 The Role of Each Stage

Here we provide the detailed observations of Section 4.2 in the main content.

Removing Stage 1: Stage 1 aims to enhance the reasoning capabilities of the models when solving medical problems. As depicted in Table 2 (2 → 3), the absence of Stage 1 leads to a notable degradation of 7.30% in overall model performance. This reduction occurs because Stage 1 serves as the foundation for Stage 2. When the model lacks strong medical reasoning capabilities, it is unlikely to recognize when external knowledge is needed or to construct effective retrieval queries for obtaining supportive information. Consequently, the potential benefits offered by subsequent training stages are substantially compromised.

Removing Stage 2: Stage 2 builds upon the groundwork established in Stage 1 to refine the generation of semantically precise and retrieval-efficient query terms, which aims to further augment the quality and utility of retrieved documents. As seen in Table 2 (1 → 3), after removing Stage 2, there is a slight decrease of 5.89% in the accuracy compared to the corresponding model trained with all three stages.

Removing Stage 3: Stage 3 is designed to further enhance the synergy between the model's retrieval and reasoning capabilities, thereby improving its overall end-to-end performance in solving medical problems. As observed in Table 2 (1 → 2), the performance of models declines by 2.88% when Stage 3 is excluded. However, due to the presence of complete Stage 1 and 2, the gap remains close to that of the complete three-stage training model and relatively low.

C.5 Environment and Hardware Configurations

The experiment utilizes the following core libraries and their respective versions: torch=2.5.1, CUDA_version=12.4, ray=2.40.0, vllm=0.7.3, verl=0.2.0.post2, transfomrers=4.49.0, datasets=3.3.2, tqdm=4.40.0, flash-attn=2.5.8, pyarrow=19.0.1, tensordict=0.5.0. Experiments are conducted using 32 NVIDIA GPUs with 96GB memory.

D Case Studies

Model	Training Method	Inference Method	MedQA-USMLE	MedQA-MCMLE	MedMCQA	RareArena-RDC	RareArena-RDS	MMLU-Med	NEJMQA	MedXpertQA	Avg.
			In-Domain (ID)					*Out-of-Domain (OOD)*			
Close-Sourced Models											
GPT-4o-mini	–	Naive	72.31	67.28	66.00	50.67	38.75	74.60	55.08	17.84	55.32
		CoT	**75.61**	68.32	69.58	48.97	42.35	80.63	57.40	20.74	57.95
		RAG	73.30	68.18	66.82	51.25	39.47	77.11	54.89	15.63	55.83
		CoT-RAG	74.45	**69.25**	**70.52**	**52.03**	**43.24**	**80.96**	**57.87**	**21.58**	**58.74**
Open-Sourced Medical-Specific Models											
MEDITRON-7B	–	CoT-RAG	48.67	44.28	46.78	37.97	21.60	50.12	33.40	16.55	37.42
UltraMedical3-8B	–	CoT-RAG	61.57	52.42	61.82	40.76	28.54	72.52	45.31	10.82	46.72
UltraMedical3.1-8B	–	CoT-RAG	66.86	58.45	65.73	43.83	32.39	75.86	50.66	12.08	50.73
MEDITRON-70B	–	CoT-RAG	60.60	55.64	56.48	75.16	48.85	70.53	65.33	18.72	56.41
Open-Sourced Medical Reasoning Models											
HuatuoGPT-o1-8B	–	Naive	62.67	62.78	65.45	45.86	36.02	70.80	51.57	13.39	51.07
		CoT	**67.84**	65.40	70.88	46.42	39.15	74.36	51.03	14.17	53.66
		RAG	64.52	63.75	66.76	48.50	38.94	72.18	50.63	13.97	52.41
		CoT-RAG†	66.97	**66.15**	**72.45**	**49.76**	**41.59**	**74.52**	**51.60**	**14.34**	**54.67**
MedS^3-8B	–	Naive	58.14	60.24	60.20	44.60	33.79	72.12	53.61	12.01	49.34
		CoT	69.11	65.23	61.35	43.79	34.08	74.36	52.17	**13.64**	51.72
		RAG	63.67	62.68	62.51	**47.83**	37.14	70.50	52.97	12.23	52.44
		CoT-RAG	71.59	66.84	64.95	45.17	34.63	75.38	52.89	11.76	52.90
		P-VS◇†	**73.51**	**69.63**	**65.47**	46.85	**37.72**	**78.75**	**55.09**	12.50	**54.94**
AlphaMed-8B	–	Naive	59.30	62.50	62.85	40.88	35.36	70.63	50.72	23.67	50.74
		CoT	63.82	64.49	65.85	41.57	37.38	**72.26**	50.93	**25.43**	52.72
		RAG	62.70	64.17	63.96	41.83	**38.22**	71.85	51.04	20.75	51.82
		CoT-RAG†	**64.06**	**64.98**	**66.43**	**43.55**	38.14	71.44	**51.48**	22.01	**52.76**
Open-Sourced Base / Instruct Models											
LLaMA3.1-8B -Instruct	–	Naive	31.16	41.45	30.02	41.85	22.16	37.12	50.41	14.90	33.63
		CoT-RAG	33.80	41.69	35.75	43.12	25.99	42.83	52.97	11.24	35.92
	SFT	CoT-RAG	61.39	62.10	63.27	44.83	35.20	63.08	49.64	12.72	49.03
	R1-Searcher∗	CoT-RAG	60.28	60.92	63.54	42.87	38.65	70.19	53.81	15.73	50.75
	ReSearch∗	CoT-RAG	62.76	66.03	66.25	46.35	38.44	71.27	53.26	14.65	52.38
	Med-R^3	Naive	67.17	62.35	71.06	52.06	41.20	74.15	56.00	12.29	54.54
		CoT	69.43	65.58	73.78	52.63	43.87	76.84	56.39	14.76	56.66
		RAG	70.09	67.56	74.80	55.00	45.52	76.25	58.10	13.94	57.66
		CoT-RAG†	**75.91**	**75.95**	**75.89**	**57.34**	**47.16**	**79.07**	**60.60**	**16.48**	**61.05**
Qwen2.5-7B	–	Naive	22.58	39.14	28.77	32.17	23.10	44.45	41.48	11.59	30.41
		CoT-RAG	27.45	37.33	28.21	30.80	24.85	48.16	45.52	10.97	31.66
	SFT	CoT-RAG	52.56	50.04	57.90	53.45	34.67	56.94	48.23	11.28	45.63
	R1-Searcher∗	CoT-RAG	56.78	49.70	58.35	53.69	33.27	66.81	52.98	12.55	48.02
	ReSearch∗	CoT-RAG	62.47	60.24	63.11	55.95	34.68	70.29	52.30	12.67	51.46
	Med-R^3	Naive	60.82	55.60	62.28	59.98	37.56	69.72	54.09	12.22	51.53
		CoT	64.48	62.28	63.74	61.05	38.09	73.73	52.57	13.85	53.72
		RAG	65.99	64.83	66.18	60.86	42.49	71.04	55.92	12.78	55.01
		CoT-RAG†	**68.64**	**67.53**	**68.97**	**63.02**	**45.76**	**75.81**	**58.54**	**14.98**	**57.91**
Qwen2.5-14B	–	Naive	50.01	50.70	42.85	43.17	26.58	71.60	45.63	11.06	42.70
		CoT-RAG	56.83	58.75	52.67	61.26	44.39	73.48	46.02	11.14	50.57
	SFT	CoT-RAG	68.85	70.22	70.15	75.27	52.82	75.81	47.08	11.54	58.97
	R1-Searcher∗	CoT-RAG	69.20	71.75	68.45	76.32	54.05	77.69	52.08	12.65	60.27
	ReSearch∗	CoT-RAG	69.52	74.05	72.20	75.67	53.54	80.25	50.65	13.10	61.12
	Med-R^3	Naive	72.98	73.40	71.78	**78.80**	55.45	78.87	55.35	13.67	62.54
		CoT	72.30	74.82	70.44	76.95	56.37	83.67	59.81	15.33	63.71
		RAG	76.58	77.75	73.28	78.36	57.11	81.49	56.64	14.01	64.40
		CoT-RAG†	**78.01**	**80.59**	**75.42**	77.94	**58.15**	**85.33**	**62.40**	**15.69**	**66.69**

Table 6: Comprehensive comparison of Med-R^3 with baselines, where we have provided the inference strategy for each competitor. Here we have selected the highest-performing one (marked with †) to represent the optimal performance of each method, as summarized in Table 1. ∗ denotes our re-implementation with the same amount of our constructed training data for a fair comparison. ◇ represents the PRM guided Vote Sum (P-VS) strategy during inference, which unlocks the full potential of MedS^3-8B (Jiang et al., 2025). The best and second best of each model are in **bold** and underlined.

Order	Model	MedQA-USMLE	MedQA-MCMLE	MedMCQA	RareArena-RDC	RareArena-RDS	MMLU-Med	NEJMQA	MedXpertQA	Avg.
		In-Domain (ID)					*Out-of-Domain (OOD)*			
Standard Pipeline										
$1 \rightarrow 2 \rightarrow 3$	Qwen2.5-7B	**68.64**	**67.53**	**68.97**	63.02	**45.76**	**75.81**	**58.54**	**14.98**	**57.91**
	LLaMA3.1-8B-Instruct	**75.91**	**75.95**	**75.89**	**57.34**	**47.16**	**79.07**	**60.60**	16.48	**61.05**
	Qwen2.5-14B	**78.01**	**80.59**	75.42	**77.94**	**58.15**	**85.33**	**62.40**	**15.69**	**66.69**
Necessity of Progressive Training										
1 & 2 & 3	Qwen2.5-7B	64.83	64.18	65.26	62.87	43.04	72.80	54.35	13.85	55.15
	LLaMA3.1-8B-Instruct	70.67	71.85	70.79	55.12	44.78	75.43	58.64	15.34	57.83
	Qwen2.5-14B	75.30	79.04	73.38	77.00	55.89	84.65	60.16	15.42	65.11
Effectiveness of Stage 1 (Reasoner Cultivation)										
$2 \rightarrow 3$	Qwen2.5-7B	62.92	60.96	63.78	61.57	40.05	71.12	54.87	14.14	53.68
	LLaMA3.1-8B-Instruct	66.85	69.36	68.95	51.11	43.80	73.62	55.84	14.85	55.55
	Qwen2.5-14B	72.69	74.81	72.81	75.33	56.95	80.42	57.04	14.33	63.05
Effectiveness of Stage 2 (Retriever Awakening)										
$1 \rightarrow 3$	Qwen2.5-7B	65.71	63.52	66.15	60.14	40.88	71.69	53.96	13.97	54.50
	LLaMA3.1-8B-Instruct	72.58	72.43	73.02	53.17	42.19	75.80	57.28	13.62	57.51
	Qwen2.5-14B	74.80	76.74	74.62	77.68	56.07	82.58	57.83	14.95	64.41
Effectiveness of Stage 3 (Dual-Process Collaboration)										
$1 \rightarrow 2$	Qwen2.5-7B	66.84	65.45	66.94	63.18	43.63	73.65	55.63	14.56	56.24
	LLaMA3.1-8B-Instruct	69.73	70.67	71.56	54.28	45.14	78.23	59.32	**16.80**	58.22
	Qwen2.5-14B	76.55	80.28	74.03	76.97	57.46	83.26	59.47	14.78	65.35
Sequential Order of Stages										
$2 \rightarrow 1 \rightarrow 3$	Qwen2.5-7B	65.98	65.78	67.52	**64.46**	42.72	74.90	55.28	14.80	56.43
	LLaMA3.1-8B-Instruct	71.49	73.07	73.60	53.26	45.98	77.44	60.09	15.93	58.86
	Qwen2.5-14B	77.45	78.57	**75.59**	76.54	57.72	82.91	60.75	15.16	65.59

Table 7: **Original scores for each benchmark of the ablation study on multiple training stages and sequential order.** Order: sequential order of Stage 1, 2, and 3. Specifically, "1 & 2 & 3" refers to a joint training configuration in which reward functions from all three stages are merged and optimized concurrently. The best and second best scores of each model are in **bold** and underlined. We evaluate these models under the inference strategy of ***CoT-RAG*** (interleaved reasoning and retrieval).

Method	MedQA-USMLE	MedQA-MCMLE	MedMCQA	RareArena-RDC	RareArena-RDS	MMLU-Med	NEJMQA	MedXpertQA	Avg.
	In-Domain (ID)					*Out-of-Domain (OOD)*			
Naive Response	22.58	39.14	28.77	32.17	23.10	44.45	41.48	11.59	30.41
Analysis of $\mathcal{R}_{reasoning} = R_{semantic} + R_{statistic} + R_{logical}$									
w/o $R_{semantic}$	66.47	67.19	67.25	**64.33**	43.68	75.04	55.36	14.86	56.77
w/o $R_{statistic}$	64.06	63.34	64.80	61.28	41.53	72.47	55.91	13.72	54.64
w/o $R_{logical}$	64.95	65.81	66.84	62.67	44.46	74.28	57.42	14.30	56.34
Analysis of $\mathcal{R}_{retrieval} = R_{quality} + R_{breadth}$									
w/o $R_{quality}$	68.18	66.07	67.35	62.46	42.79	73.82	54.73	14.67	56.26
w/o $R_{breadth}$	66.35	64.84	66.92	61.70	43.58	73.40	55.34	14.06	55.77
Med-R^3	**68.64**	**67.53**	**68.97**	63.02	**45.76**	**75.81**	**58.54**	**14.98**	**57.91**

Table 8: **Original scores for each benchmark of the ablation study on reward design. w/o**: *removing* the reward component from the original metric. The best and second best scores are marked in **bold** and underlined. We employ ***Qwen2.5-7B*** for experimental analysis, and evaluate the model under the inference strategy of ***CoT-RAG*** (interleaved reasoning and retrieval).

Question

A 41-year-old construction worker presents to the office complaining of a progressively worsening breathlessness for the last 2 months. He has no other complaints. His medical history is significant for hypertension being treated with lisinopril-hydrochlorothiazide and gastroesophageal reflux disease being treated with pantoprazole. He has a 30-pack-year smoking history and drinks alcohol on the weekends. He works mainly with insulation and drywall placing. His temperature is 37.0°C (98.6°F), the blood pressure is 144/78 mm Hg, the pulse is 72/min, and the respirations are 10/min. Upon further questioning about his employment, the patient admits that he does not regularly use a mask or other protective devices at work. Given his occupational exposure, smoking history, and symptoms, what malignancy is this patient most likely at risk for?

Ground Truth *Lung cancer*

Answer

LLaMA3.1-8B-Instruct

<think> Note that the patient is a 41-year-old construction worker who presented with a progressively worsening breathlessness for the last 2 months, which could be a symptom of many respiratory issues. The patient has a significant smoking history, which puts him at risk for lung cancer as well as other respiratory cancers. His presenting symptom of progressive breathlessness over the past two months raises concern for a pulmonary or pleural-based malignancy. Asbestos exposure is strongly linked to both lung cancer and mesothelioma. While lung cancer is more common overall, mesothelioma is more specifically associated with asbestos exposure. The latency period for mesothelioma is long (often decades), and symptoms such as dyspnea, chest pain, and cough may appear only when the disease is advanced. Considering the occupational exposure pattern, lack of protective measures, and clinical presentation, "mesothelioma" should be considered a likely concern in this patient. </think><answer> Mesothelioma </answer>

LLaMA3.1-8B-Instruct + Med-R^3

<think> Given the patient's occupational exposure to insulation and drywall materials, his significant smoking history, and the progressive breathlessness symptom, it is important to consider malignancies linked to these risk factors. Insulation work often involves exposure to asbestos, a known carcinogen associated with mesothelioma and lung cancer. Additionally, smoking further increases the risk of lung cancer. Breathlessness may suggest lung involvement or pleural disease. <search> occupational exposure asbestos-related cancers </search><document> Epidemiological studies have consistently demonstrated a strong association between occupational asbestos exposure and the development of lung cancer and mesothelioma. \n Studies have shown that workers in industries such as construction, shipbuilding, and asbestos mining are at a significantly higher risk of developing asbestos-related cancers. The latency period between initial exposure and the onset of cancer can range from 10 to 50 years, making early detection and prevention crucial. \n Meta-analysis of 25 studies involving over 100,000 workers exposed to asbestos revealed a significant increase in the risk of lung cancer. The standardized mortality ratio (SMR) for lung cancer was 2.8 (95% CI: 2.5-3.1). These findings underscore the substantial health risks associated with occupational asbestos exposure. </document> The patient is at an increased risk for lung cancer and mesothelioma due to his occupational exposure to asbestos, significant smoking history, and the presence of progressive breathlessness. Asbestos exposure is strongly associated with both malignancies, particularly in individuals working in construction, insulation, and drywall installation without protective equipment. Smoking synergistically increases the risk of lung cancer in individuals exposed to asbestos. Given that the patient has been exposed to both risk factors—smoking and asbestos—the most likely malignancy he is at risk for is lung cancer. </think><answer> Lung cancer </answer>

Figure 8: Case study of Med-R^3.

E Prompts

Here we present the prompts used throughout our pipeline in Med-R^3 . Only the English version is presented due to LaTeX compilation issues with non-English languages.

Prompt: Content Suitability Judgment

Please judge whether the following multiple-choice question is suitable for conversion into an open-ended question. The question to be converted must meet the following conditions:

1. After removing the options, the question itself remains valid, and the answer is unique and correct.
2. There is no ambiguity in the question and answer.
3. The question must have a unique optimal answer, not a range or a vague value.
4. Questions with negative options, such as selecting the option that does not meet the conditions or the "least likely" option, are not suitable.
5. In other cases, please use your logical judgment to make a decision.

 # Question
{question}

 # Correct Answer
{answer}

 # Misleading Options
{misleading_options}

Output Format:

```json
{
    "unique": True,
    "reason": "Because..."
}
'''
```

Prompt: Reasoning Complexity Filtering

You are an expert in the medical field. You will be given a question, a student's answer, the correct and unique answer to the question, and other misleading options. Please compare the student's answer with the correct answer and analyze whether the student's answer is correct. In the student's answer, the `<think>...</think>` tag wraps the thinking process, and the `<answer>...</answer>` tag wraps the final answer. During the model's reasoning process, uncertain parts are encapsulated using the `<search>`*search query*`</search>` tags to facilitate subsequent information retrieval. Once relevant content is retrieved, the model continues its reasoning based on the retrieved information. It is important to note that the content within `<document>` *search results* `</document>` represents externally retrieved information and is not generated by the model itself. Therefore, this content should not be used to assess the model's logical reasoning ability. Instead, it should be used solely to evaluate the model's capability to process and integrate external information.

Please judge the student's thinking process and answer based on the correct answer, rate it from 1 to 5 points, and explain the reason.

5 Point Answer Criteria:

1. The thinking process is logical and seamless, and the reasoning process is specific and clear.
2. The final answer is consistent with the correct answer, allowing synonyms, abbreviations, etc. of the correct answer, but cannot contain incorrect options.

3-4 Point Answer Criteria:

1. The thinking process is reasonable.
2. The final answer is consistent with the correct answer, allowing some supplements, as long as they do not conflict

with the correct answer, and the correct answer is the main one, not other misleading options.

1-2 Point Answer Criteria:

1. The thinking process is not clear.
2. The final answer is inconsistent with the correct answer or contains incorrect options.
3. Contains garbled characters, format errors, disorder, and irrelevant information.

 # Question
{question}

 # Student's Answer
{answer}

 # Correct Answer
{correct_answer}

 # Misleading Options
{misleading_options}

Output Format:

```json
{
    "score": xxx,
    "reason": "..."
}
'''
```

Prompt: Open-Ended Standardization

You are an expert in question reformulation within the medical field. Please convert the following multiple-choice question into an open-ended question. Please try to keep the content of the question unchanged as much as possible, and modify the last question into an open-ended inquiry, that is, modify the original "Which of the following is", "The most likely option is", etc. The modified question should also be semantically smooth and unambiguous, and the answer to the question should be consistent with the correct answer to the original question. The language used in the modified question should be consistent with the original question.

 # Original Question
{question}

 # Answer
{answer}

Output Format:

```json
{
    "question": "..."
}
'''
```

Prompt: Reasoning Process Generation (with Retrieval, for training data construction, rollout in Med-R^3, and inference mode of *CoT-RAG*)

You are a medical expert. Given a question, you should answer it by first thinking about the reasoning process in the mind and then providing the final answer. Please answer the question in the format of `<think>`...`</think><answer>`...`</answer>`. That is, `<think>`*Here is the reasoning process*`</think><answer>`*answer*`</answer>`. You should perform thinking with decomposing, reflecting, brainstorming, verifying, refining, and revising. Besides, you can perform searching for uncertain knowledge if necessary with the format of `<search>`*search query*`</search>` during your thinking process. Then, the search system will provide you with the retrieval information with the format of `<document>` *search results* `</document>`. The answer needs to summarize the reasoning process and give the final answer. You are required to continue your reasoning and response in conjunction with the existing answer and retrieved information, ensuring that the subsequent answers you generate maintain coherence with the previously generated

answers.

 # Question
{reformulated_question}

 # Existing Answer and Retrieved Information
{existing_answer} [optional]

Prompt: Reasoning Process Generation (without Retrieval, for inference mode of *CoT*)

You are a medical expert. Given a question, you should answer it by first thinking about the reasoning process in the mind and then providing the final answer. Please answer the question in the format of `<think>...</think><answer>...</answer>`. That is, `<think>`*Here is the reasoning process*`</think><answer>`*answer*`</answer>`. You should perform thinking with decomposing, reflecting, brainstorming, verifying, refining, and revising. The answer needs to summarize the reasoning process and give the final answer.

 # Question
{reformulated_question}

Prompt: Answer Generation (without Retrieval, for inference mode of *Naive*)

You are a medical expert, please provide answer for the following question.

 # Question
{reformulated_question}

Prompt: Answer Generation (with Retrieval, for inference mode of *RAG*)

You are a medical expert. Given the following question, please consult the retrieved documents, identify key information that are directly related to the question, and provide the answer.

 # Question
{reformulated_question}

 # Retrieved Documents
{retrieved_documents}

Prompt: Medical Knowledge Graph Extraction

You are a medical expert. Given a reasoning process for solving a medical problem in the format of `<think>...</think>` and `<answer>...</answer>`. The content within `<think>...</think>` demonstrates the thought process and may use `<search>`*search query*`</search>` to mark uncertain knowledge that requires searching. The search system provides relevant information in the format of `<document>` *search results* `</document>`. Your task is to extract important medical concepts, relationships, and attributes from the given reasoning process and represent them in the knowledge graph format.

 # Reasoning process
{reasoning_reference}

Output Format:

```
```json
[[
    "entity1", "relationship",
    "entity2", "if_retrieval"]
], ...]
'''
```

Definitions:

1. **entity1, relationship, entity2**: Clearly extract entities or relationships from the natural language (e.g., "patient", "has symptom", "fever").
2. **if_retrieval**: A boolean value (`1`/`0`) indicating whether the entity or relationship is retrieved from the search results within `<document>...</document>`.

Processing rules:
```

1. ***Overall***: Extract only entities and relationships related to solving medical problems, ignoring irrelevant background information. The knowledge graph should be concise and clear, avoiding redundancy, and based on facts, avoiding subjective speculation.

2. ***Entity Extraction***: Must be specific medical concepts (diseases, symptoms, drugs, etc.), excluding vague descriptions (such as "some conditions" or "related factors"). Synonymous expressions should be merged (e.g., "myocardial infarction" instead of "heart attack").

3. ***Relationship Definition***: Use verb phrases (cause, inhibit, lead to, accompany, etc.) and must have a clear directionality (A→B or B→A should conform to medical logic). Negative relationships should be explicitly marked (e.g., "does not cause", "rule out").

4. ***Retrieval Marking Determination***: If an entity or relationship is directly from the `<document>` *search results* `</document>`, mark "`if_retrieval=1`"; otherwise, "`if_retrieval=0`". The search keywords themselves should not be considered as entities or relationships.

5. ***Special Handling***: Retain hypothetical relationships in the reasoning process (marked as [hypothesis]), convert time relationships to medical temporal expressions (acute/chronic/ongoing period, etc.), and quantify probabilistic conclusions as (high/medium/low) risk levels.

Prompt: Evidence Quality Judgment

You are an expert in evidence quality annotation within the medical field. There are 6 quality levels of evidence, ranging from the highest to the lowest as follows: Meta-Analyses and Systematic Reviews, Randomized Controlled Trials, Cohort Studies, Case-Control Studies, Individual Case Series or Case Reports, Background Information or Expert Opinion. Please classify the following evidence document based on its structure and characteristics, providing only the names of the levels, without any additional description:

 # Evidence
{retrieved_document}

Prompt: Answer Correctness Judgment

You are an expert in the medical field. You will be given a question, a student's answer, and the correct answer to the question. Please compare the student's answer with the correct answer and analyze whether the student's answer is correct. In the student's answer, the `<think>...</think>` tag wraps the thinking process, and the `<answer>...</answer>` tag wraps the final answer. Please judge the student's answer encapsulated within the `<answer>...</answer>` tag based on the correct answer, rate it from 0 to 2 points, and explain the reason.

2 Point Answer Criteria:
The final answer is consistent with the correct answer, allowing synonyms, abbreviations, etc.

1 Point Answer Criteria:
The final answer is consistent with the correct answer, allowing some supplements, as long as they do not conflict with the correct answer.

0 Point Answer Criteria:
The final answer is inconsistent with the correct answer, or contains garbled characters, format errors, disorder, and irrelevant information.

 # Question
{question}

 # Student's Answer
{answer}

 # Correct Answer
{correct_answer}

Output Format:

```json
{
    "score": xxx,
    "reason": "..."
}
'''
```