ReasonMed: A 370K Multi-Agent Generated Dataset for Advancing Medical Reasoning

Anonymous ACL submission

Abstract

Though reasoning-based large language models (LLMs) have excelled in mathematics and programming, their capabilities in knowledge-intensive medical question answering remain underexplored. To address this, we introduce ReasonMed, the largest medical reasoning dataset, comprising 370k high-quality examples distilled from 1.7 million initial reasoning paths generated by various LLMs. ReasonMed is constructed through a *multi-agent verification and refinement process*, where we design an *Error Refiner* to enhance the reasoning paths by identifying and correcting error-prone steps flagged by a verifier. Leveraging ReasonMed, we systematically investigate best practices for training medical reasoning models and find that combining detailed Chain-of-Thought (CoT) reasoning with concise answer summaries yields the most effective fine-tuning strategy. Based on this strategy, we train ReasonMed-7B, which sets a new benchmark for sub-10B models, outperforming the prior best by 4.17% and even exceeding LLaMA3.1-70B on PubMedQA by 4.60%.[1]

1 Introduction

Recent reasoning-based large language models (LLMs), such as Deepseek-R1 (DeepSeek-AI, 2025) and QwQ (Team, 2025), have garnered significant attention due to their remarkable capabilities in logical reasoning (Liu et al., 2025), mathematics (Ahn et al., 2024), and programming (OpenAI et al., 2025) tasks.

Despite their effectiveness, LLMs encounter notable challenges in the medical domain. First, the inherently knowledge-intensive nature of medicine demands large volumes of high-quality, accurately curated data for reliable reasoning. However, existing medical reasoning datasets, such as medical-o1-reasoning-SFT and Medical-R1-Distill-Data (Chen et al., 2024), are limited in size and typically derived from a single teacher model, restricting their knowledge coverage. Furthermore, current studies lack a systematic analysis of the trade-offs between resource-intensive, multi-step CoT reasoning (Wei et al., 2023) and more compact, summary-based approaches. It remains an open question whether the added cost of explicit reasoning justifies its performance benefits over more efficient summarization strategies in medical QA systems.

To tackle these challenges, we present ReasonMed, a large-scale medical reasoning dataset comprising 370k rigorously verified examples, which is an order of magnitude larger than prior datasets (Chen et al., 2024). Sampled from multiple competitive LLMs, ReasonMed integrates diverse medical insights, enhancing its depth and coverage. Each example includes both detailed multi-step CoT reasoning and a concise answer summary, facilitating analysis of effective reasoning patterns in the medical domain.

Dataset scale plays a crucial role in enhancing model performance. To this end, we adopt a large-scale, high-quality data generation paradigm using a multi-agent system (MAS). We first aggregate approximately 195k questions (excluding test splits) from for established benchmarks: MedQA (Jin et al., 2020), MMLU (Hendrycks et al., 2021), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022). Our MAS combines three competitive LLMs, two general-purpose models (Qwen-2.5-72B (Team, 2024) and DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025)) and one medical-specific model (HuatuoGPT-o1-70B (Chen et al., 2024)). By manipulating sampling hyperparameters (e.g., temperature, top-p) across agents, we generate around 1.75 million diverse, multi-step reasoning paths. This combination of scale and methodological rigor is designed to boost data quality and, consequently, improve model performance on complex clinical

[1]All code and resources will be released publicly.

QA tasks.

Beyond dataset size, training efficacy is highly sensitive to data quality. Prior work (Muennighoff et al., 2025) shows that excellent performance is attainable with as few as $1,000$ high-quality examples. To reach comparable precision in medical QA, we devise a rigorous quality control pipeline that validates every reasoning chain for *answer correctness*, *logical coherence*, and *medical factuality*. Through the pipeline, questions are categorized by validation pass rate into three tiers: *easy* (≥ 5 correct paths), *medium* (2-4 correct paths), and *difficult* (< 2 correct paths). For easy questions, the two top-ranked reasoning paths verified by a quality ranker are retained. For medium questions, because subtle yet frequent errors persist, an *error refiner*, driven by verifier logs and powered by GPT-4o-mini, is applied to revise and expand the selected reasoning paths. For difficult questions, we directly employ GPT-o1 with a structured multi-step process to generate correct reasoning paths. Through this multi-stage refinement process, we produce a polished dataset of 370 K high-quality medical reasoning samples.

In addition to generating high-quality reasoning data, we also investigate the impact of various reasoning training strategies on model performance. Specifically, we compare fine-tuning approaches including traditional chain-of-thought (CoT), summary-based responses, and a hybrid CoT-summary method. Using `lm_eval` framework (Gao et al., 2024) for rigorous evaluation, we identify the most effective strategies for improving medical LLMs on complex questions. Results show that the hybrid approach yields the highest accuracy, while summary-only responses offer competitive performance with lower computational cost, highlighting the potential for strategy selection based on application needs.

Our main contributions are fourfold:

- We release the largest open-source medical reasoning dataset, comprising around 1.29 million validated paths, refined to 370k high-quality examples via targeted optimization.
- We construct a multi-agent framework for generating, filtering, and optimizing reasoning paths. Evaluated by GPT-4o on randomly sampled subsets of $1,000$ and $3,000$ entries, our ReasonMed dataset demonstrates superior overall quality compared to data generated by GPT-4o and DeepSeek-R1.
- We present the first systematic evaluation of explicit reasoning in knowledge-intensive medical QA, using a consistent dataset to comprehensively assess performance, computational efficiency, and accuracy.
- The trained ReasonMed-7B model achieves state-of-the-art performance among sub-10B models and surpasses several larger counterparts on medical QA benchmarks.

2 Related Work

Multi-Agent-based Data Curation. The use of multi-agent frameworks has emerged as a robust approach to dataset generation and optimization across various domains. These systems often employ specialized agents collaboratively performing tasks analogous to human team problem-solving (Hong et al., 2023). Recent works such as DialogueAgents (Li et al., 2025) leverage specialized agents including scriptwriters, synthesizers, and critics to generate high-quality, diverse dialogue datasets. In the programming domain, AgentCoder (Huang et al., 2024) uses agents such as programmers, test designers, and test executors, significantly enhancing the robustness of generated data through iterative agent-driven feedback. BOLT (Pang et al., 2025) integrates multi-agent frameworks with large language models (LLMs) to produce long-chain reasoning data, further highlighting the efficacy of this approach in creating structured, reasoning-intensive datasets. Unlike previous multi-agent applications, our framework specifically targets medical reasoning datasets, employing specialized medical and general-purpose language models to generate, validate, and refine high-quality reasoning paths, explicitly tailored for medical QA scenarios.

Medical Reasoning Dataset & Model. Recent studies highlight the efficacy of chain-of-thought (CoT) prompting in improving model performance on medical QA benchmarks (Wei et al., 2022; Liévin et al., 2023). Models employing adaptive reasoning, such as medical language agents, have been introduced to systematically address complex clinical tasks (Dutta and Hsiao, 2024). Furthermore, multi-agent systems, employing specialized medical reasoning agents, collaboratively synthesize clinical insights, thus enhancing decision-making reliability and interpretability (Zuo et al., 2025). HuatuoGPT (Chen et al., 2024) further exemplifies the integration of comprehensive medical

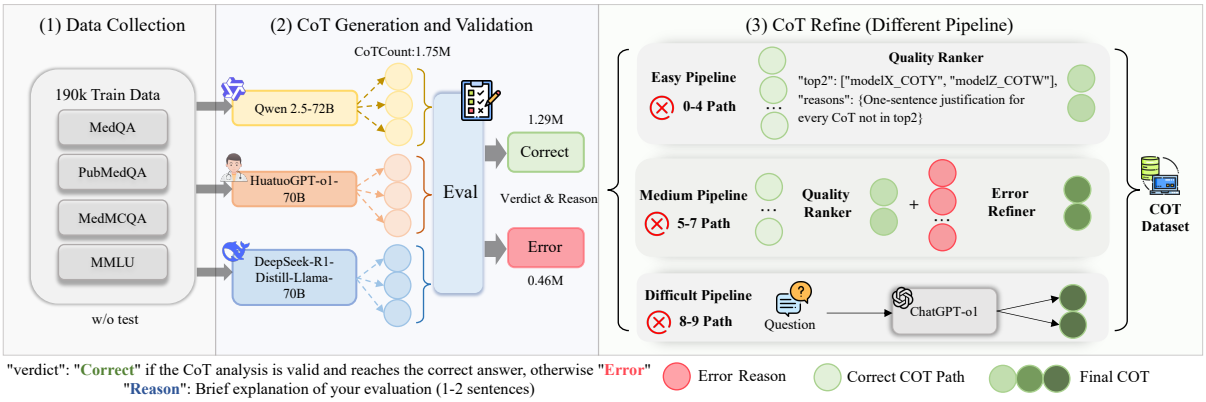

Figure 1: (1) show composition of the dataset. (2) present the Multi-Agent System for generating and validating Complex CoT. (3) outline strategy schemes (Easy/Medium/Difficult Pipeline) based on CoT validation counts. For 0-4 errors, select top two CoTs using the Quality Ranker. For 5-7 errors, optimize the top two CoTs with GPT-4o-mini, addressing identified weak points. For 8-9 errors, generate high-quality answers using GPT-o1.

knowledge and multi-step reasoning into large language models. However, existing datasets often lack rigorous verification processes and structured optimization strategies tailored to medical QA complexity. Our work uniquely addresses this gap by employing a rigorous, multi-stage optimization and verification pipeline, systematically evaluating and refining multi-step reasoning paths to significantly enhance the quality and applicability of the resulting medical reasoning dataset.

LLM-as-a-Judge. Employing large language models as evaluators (LLM-as-a-Judge) has become increasingly prevalent, providing scalable and consistent assessment frameworks across various domains (Gu et al., 2025). Notably, in medical QA tasks, LLM evaluators have demonstrated enhanced evaluation consistency and accuracy (Krolik et al., 2024; Zhao et al., 2024). LLM-based evaluators iteratively assess and refine reasoning steps, guiding models toward correct and logically coherent paths (Qin et al., 2024). Approaches such as QuRating (Tang et al., 2024) have underscored the potential for systematic selection of high-quality training data using LLM evaluators. In contrast to existing studies, our approach evaluates the language modelgenerated CoT reasoning paths for correctness and potential factual errors, and additionally outputs the error reasons for flawed paths to facilitate subsequent optimization. We also developed a Score Evaluator to offer an assessment framework comparing reasoning paths before and after optimization and datasets quality.

3 Multi-Agent Reasoning Pipeline

3.1 Dataset Composition

In this section, we present the composition of the dataset used for the Multi-Agent Reasoning Pipeline, along with an analysis of the dataset's structure and the benchmarks involved. The dataset consists of various medical question-answering datasets. Table 1 shows a summary of the dataset composition:

Dataset Composition	Count
MedQA (train/dev)	10178/1272
MedMCQA (train)	182822
PubMedQA (train/val)	450/50
MMLU	
Anatomy (dev/val)	5/14
Clinical Knowledge (dev/val)	5/29
College Biology (dev/val)	5/16
College Medicine (dev/val)	5/22
Medical Genetics (dev/val)	5/11
Professional Medicine (dev/val)	5/31
Total Count	**194925**

Table 1: Summary of ReasonMed Question Count Composition.

3.2 Multi-Agent System for Complex CoT Generation

We employ a multi-agent framework—comprising Qwen-2.5-72B, HuatuoGPT-o1-70B, and DeepSeek-R1-Distill-Llama-70B—to generate

1.755 million reasoning paths. Each model produces three CoT trajectories at different temperatures (0.7, 0.9, and 1.0). We then assemble the complex CoTs by following these steps:

(i) Rewrite the question.

(ii) Highlighting key clinical details and background information.

(iii) Evaluate each answer choice and discussing supporting evidence and potential traps.

(iv) Systematically eliminate choices inconsistent with the clinical context.

(v) Reassess each option, eliminating inconsistencies.

(vi) Conclude with a final answer, supported by a concise explanation of the reasoning.

In Fig 2, we present a pairwise comparison among DeepSeek-R1-Distill-Llama-70B, HuatuoGPT-o1-70B, and Qwen2.5-72B on the Medical QA task. Specifically, we compare the number of questions correctly answered by each model individually. The results reveal that different models exhibit distinct strengths across various medical knowledge domains.The observed differences in knowledge domains across models highlight the necessity of a multi-agent system that integrates diverse model outputs.

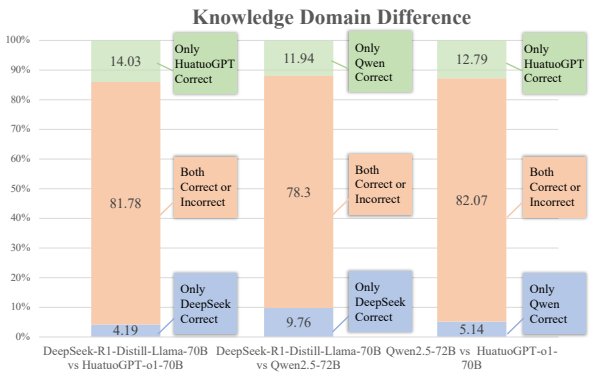

Figure 2: Knowledge domain differences among DeepSeek-R1-Distill-Llama-70B, HuatuoGPT-o1-70B and Qwen2.5-72B.

3.3 Component Design

This section provides an overview of the components developed in this paper and their respective functions. (2)-(6) of Fig 3 visualize the structure and workflow of each component.

Verifier: This component constructs a verifier (based on Qwen2.5-72B) to validate the correctness of CoT paths generated by the Multi-Agent system. The model not only checks whether the answer is correct or incorrect, but also evaluates whether the key clinical factors have been accurately identified, whether all answer choices have been analyzed, and whether there are any factual errors in the medical knowledge. The model outputs a JSON object with two keys: one indicating the verdict (Correct or Error), and the other providing the reason for the error. For example, "The CoT analysis contains inaccuracies regarding vasopressin's role in glycogenolysis and incorrectly dismisses oxytocin without full consideration of its potential regulatory effects.". Fig 4 presents a bar chart showing the number of correct versus incorrect reasoning paths—after Verifier validation—for each model and CoT configuration across the nine generated paths. DeepSeek-R1-Distill-Llama-70B achieves the highest overall accuracy; Qwen-2.5-72B retains the most correct paths at a temperature of 0.9, while the optimal temperature for the other two models is 0.7.

Response Summarizer: To construct a response with reasoning similar to o1 answers, we use GPT-4o-mini as a summarization assistant. The model generates a summary for each complex CoT, which represents a step-by-step reasoning process. This summary is presented as the final output to the user, focusing on the reasoning aspect of the response.

Quality Ranker: Balancing dataset size and quality is crucial. Among the many correct CoT paths, we aim to select the two most optimal ones for subsequent training. The Quality Ranker, based on Qwen2.5-72B, plays a critical role here. The model reads the correct CoT paths and outputs the top two, such as "top2": ["modelX_COTY", "modelZ_COTW"], along with the rationale for excluding the other options. Initially, we considered using a Score Evaluator to rate each CoT, but this approach was challenging due to cases where multiple CoTs might have identical scores, making it difficult to select the best. Therefore, we opted for directly outputting the two best paths by their CoT names. Fig 5 shows the distribution of the top two CoT paths selected by the Quality Ranker in both Easy Pipeline and Medium Pipeline, illustrating the sampling proportions across different models and temperature settings.

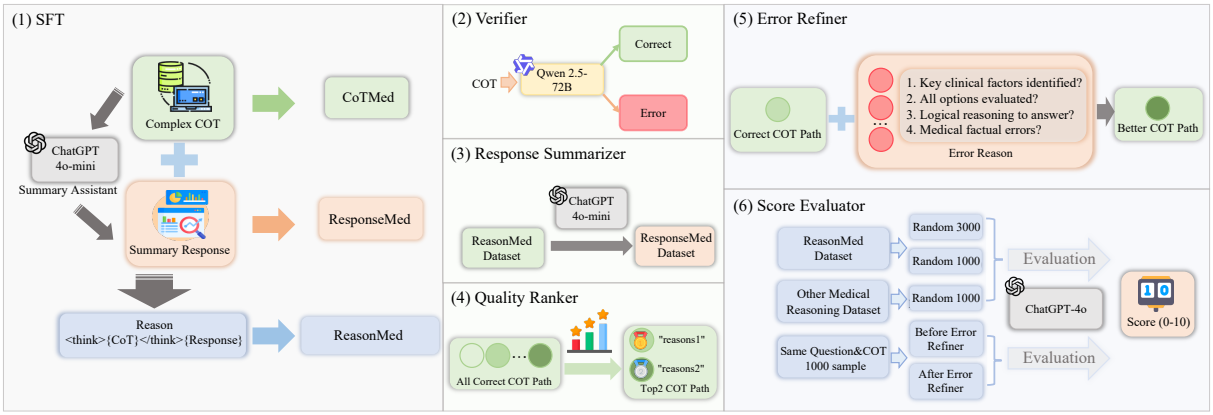

Figure 3: (1) Shows an example of SFT applied at different scales. (2) to (6) represent the components used to build the entire pipeline for our dataset.

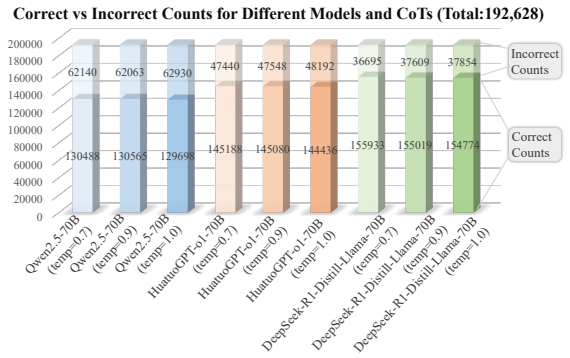

Figure 4: Bar chart illustrating the correct and incorrect counts for each model and CoT configuration across 9 generated paths in a Multi-Agent System, totaling 192,628.

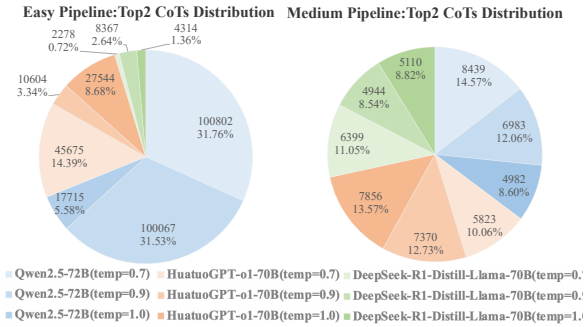

Figure 5: Distribution of the top two CoT paths selected by the Quality Ranker in Easy Pipeline and Medium Pipeline, showing sampling proportions across models and temperature settings.

Error Refiner: This component handles questions of moderate difficulty. Using the Quality Ranker, it first selects the two most optimal reasoning paths (if only two chains of thought are correct, they are chosen by default), and then performs a secondary optimization. Its design also includes storing the models error reasons during the verification stage and leveraging a stronger model to supplement and address those weak pointsan approach that effectively corrects the models error-prone knowledge.

Score Evaluator: This component utilizes the GPT-4o API to score the dataset quality on a scale from 0 to 10. We conducted two main experiments: the first compared the scores of the same question before and after CoT optimization to validate the effectiveness of the Error Refiner; the second involved comparing our final ReasonMed with other open-source medical reasoning datasets through random sampling to assess the effectiveness of our Multi-Agent approach.

3.4 ReasonMed Build Pipeline

Based on the number of errors detected in the reasoning paths, three distinct pipelines were created to process CoTs at varying levels of difficulty:

Easy Pipeline (Error 0-4) : This pipeline handles paths with few errors (0-4), which are relatively easy for the model to answer correctly. Here, we use Quailty Ranker to rank the correct paths, selecting the top two from the 5-9 correct options. Additionally, the model provides brief explanations as to why it did not choose other CoT paths.

Medium Pipeline (Error 5-7) : For paths with moderate errors (5-7), we assume that the model has partial knowledge but may miss certain fine-grained details. Thus, the top two CoT paths are selected using the Quality Ranker, and then refined using the Error Refiner based on the pitfalls provided by the Verifier, focusing on correcting those

errors to enhance the original correct reasoning chains.

Difficult Pipeline (Error 8-9) : For difficult questions with significant errors (8-9), the GPT-4o model may not be sufficient to correct the mistakes. Therefore, we use GPT-o1 to optimize these paths. For paths that are entirely incorrect, GPT-o1 generates high-quality CoTs from scratch, following the six-step reasoning process.

Lastly, Fig 6 presents the different pipeline quantity statistics, showing the distribution of paths handled by Easy, Medium, and Difficult Pipeline.

Figure 6: Different Pipeline Quantity Statistics.

By analyzing the number of correct paths validated by the Verifier, we can approximate each questions difficulty. Accordingly, we designed three distinct pipelines to tackle problems of varying complexity, systematically correcting errors in complex CoTs and refining the original dataset to strike an optimal balance between scale and quality.

4 Multiscale Supervised FineTuning

To assess the impact of explicit reasoning supervision on a downstream medical QA task, we propose a multiscale fine-tuning strategy leveraging three variants of our high-quality dataset. These variants are based on different granularities of reasoning, as outlined below:

- **CoT:** A complex chain of thought consisting of six reasoning steps,
- **Response:** A concise response generated by a Response Summarizer from the CoT,
- **Reason:** A combination of the complex CoT and its corresponding summarized response.

4.1 Data Preparation

Leveraging the 370 K ReasonMed introduced in Section 3, we employ a Response Summarizer to condense each chain-of-thought into a succinct answer explanation. For every question q and its corresponding CoT path $Multi - step = [step_1, \ldots, step_6]$, we generate the following instances:

- **CoT instance:**

$$[\,q; step_1, step_2, \ldots, step_6\,] \;\mapsto\; \texttt{CoT}.$$

- **Response instance:**

$$\text{Response Summarizer}(\texttt{CoT}) \;\mapsto\; \texttt{Response},$$

- **Reason instance:**

$$\texttt{<think>\{CoT\}</think>Response} \;\mapsto\; \texttt{Reason}.$$

The CoT, Response, and Reason instances are designed to encapsulate different levels of reasoning and summarization, providing a different scale of data for training.

4.2 Fine-Tuning and Training

We fine-tuned the open-source Qwen2.5-7B model using three different fine-tuning regimes, with each regime corresponding to a different data scale. Specifically, we utilized LlamaFactory to perform 3 epochs of supervised fine-tuning on the following datasets:

- **CoTMed-7B:** Fine-tuned with the CoT instances, focusing on reproducing the reasoning trace and generating the final answer.
- **ResponseMed-7B:** Fine-tuned with the Response instances, where the model is trained to generate concise summaries of the reasoning path.
- **ReasonMed-7B:** Fine-tuned with the Reason instances, combining detailed reasoning with summarized feedback.

Fig 3 (1) illustrates the SFT process. For evaluation, we used the lm_eval framework to analyze the performance of these models on benchmark tasks, examining whether multi-step reasoning could enhance the model's ability to perform medical QA. We also trained models with fewer epochs, including a variant trained for only one epoch, to assess performance differences and investigate the effect of fewer training steps. The results of these experiments will be discussed in detail in the experimental section.

4.3 Training Details

We performed full-model fine-tuning of the Qwen2.5-7B checkpoint using the LLaMA-Factory framework on a 16 x H20 GPU cluster. The ResponseMed configuration completed in approximately 9 hours, whereas CoTMed and ReasonMed required roughly 25 hours and 28 hours, respectively.

5 Experiments

5.1 Dataset Quality Evaluation

Medium Pipeline Validity Verification: To evaluate the effectiveness of the Medium Pipeline, we sampled 1,000 questions + CoT and used the Score Evaluator to assess the quality of answers both before and after applying the Medium Pipeline (GPT-4o-mini corrections). The results show a significant improvement, with an average score increase of 0.8 points post-optimization. The specific scores are as follows:

Dataset	Samples	Avg. Score
Medium Pipeline (pre-opt)	1,000	7.37
Medium Pipeline (post-opt)	1,000	8.17

Table 2: Score Evaluator results for Medium Pipeline validity.

Comparison with Open-Source Datasets: We compared the ReasonMed with two publicly open-source medical reasoning corpora: `medical-o1-reasoning-SFT` and `Medical-R1-Distill-Data`. For a fair comparison, we sampled 1,000 instances from each of these datasets and extended the ReasonMed with an additional 3,000 samples. The results demonstrate that the ReasonMed outperforms both baselines, achieving an average score of 8.45 for the 1,000 sample subset and 8.50 for the 3,000 sample subset. This represents an improvement of 3.9% and 5.9% over the other datasets, respectively.

Dataset	Samples	Avg. Score
medical-o1-reasoning-SFT	1,000	8.03
Medical-R1-Distill-Data	1,000	8.18
ReasonMed	1,000	8.45
ReasonMed	3,000	8.50

Table 3: Score Evaluator results for comparison with other datasets.

5.2 Multiscale Supervised Fine-Tuning

In this section, we present a comprehensive analysis of the experimental results obtained by fine-tuning the Qwen2.5-7B model using our proposed multiscale supervised fine-tuning (SFT) strategy. Performance comparisons across various medical question-answering (QA) benchmarks, including MedQA, MedMCQA, PubMedQA, and MMLU, are detailed in Table 7. Our results demonstrate the effectiveness of incorporating explicit reasoning supervision at multiple granularities:

CoTMed-7B consistently outperforms baseline models across most benchmarks, achieving notably higher scores in MedQA (66.3%), MedMCQA (64.7%), and PubMedQA (80.0%). This indicates that fine-tuning on complex reasoning chains substantially enhances the model's capacity to perform medical reasoning tasks.

ResponseMed-7B focusing solely on generating concise summaries of reasoning, achieved competitive results, with notable performance on MedQA (67.5%) but slightly lower overall accuracy (67.0%) compared to CoTMed-7B (69.1%). This suggests that while response summarization captures key information effectively, it may miss nuanced reasoning steps critical for more complex questions.

ReasonMed-7B which combines detailed reasoning chains and concise summaries, yielded the highest total accuracy (69.6%), particularly excelling in MedMCQA (65.1%) and PubMedQA (82.0%). This hybrid approach appears to effectively leverage the strengths of both granularities, achieving balanced and robust performance across diverse question types.

To explore the impact of training duration, we

	MedQA	MedMCQA (val)	PubMedQA	MMLU						Total Acc	Avg. token
				Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine		
Dataset_Count	1273	4183	1000	135	265	144	173	100	272	-	-
BioMistral-7B	45.6 ±1.4	41.5 ±0.8	71.0 ±2.0	76.3 ±3.7	63.0 ±3.0	62.5 ±4.1	53.8 ±3.8	67.0 ±4.7	53.3 ±3.0	48.9	60.1
Llama3-OpenBioLLM-8B	57.9 ±1.4	57.7 ±0.8	76.0 ±6.1	68.9 ±4.0	77.7 ±2.6	83.3 ±3.1	69.4 ±3.5	83.0 ±3.8	79.0 ±2.5	62.9	75.1
Llama-3-8B-UltraMedical	63.2 ±1.4	57.7 ±0.8	78.0 ±5.9	67.4 ±4.1	74.3 ±2.7	75.7 ±3.6	61.9 ±3.7	73.0 ±4.5	78.7 ±2.5	63.5	5177.7
Mistral-7B-Instruct-v0.3	52.2 ±1.4	48.2 ±0.8	82.0 ±5.5	59.3 ±4.2	69.4 ±2.8	72.9 ±3.7	56.7 ±3.8	70.0 ±4.6	66.5 ±2.9	55.9	111.8
Yi-1.5-9B-Chatbot	49.8 ±1.4	47.0 ±0.8	69.0 ±2.1	67.5 ±3.8	63.9 ±2.8	70.3 ±3.8	51.2 ±4.0	68.8 ±4.5	66.7 ±3.1	52.9	162.2
HuatuoGPT-o1-7B	**68.4 ±1.3**	57.5 ±0.8	74.0 ±2.0	71.9 ±3.9	78.5 ±2.5	**88.2 ±2.7**	67.6 ±3.6	80.0 ±4.0	77.6 ±2.5	64.4	446.0
HuatuoGPT-o1-8B	65.4 ±1.3	61.0 ±0.8	74.6 ±2.0	69.6 ±4.0	77.7 ±2.6	81.3 ±3.3	69.9 ±3.5	78.0 ±4.2	71.0 ±2.8	65.5	468.9
ResponseMed-7B (1epoch)	62.2 ±1.4	57.6 ±0.8	**84.0 ±5.2**	75.6 ±3.7	77.7 ±2.6	81.3 ±3.3	69.9 ±3.5	87.0 ±3.4	76.8 ±2.6	64.8	-
CoTMed-7B(1epoch)	64.3 ±1.3	62.4 ±0.8	82.0 ±5.5	**77.0 ±3.6**	**80.8 ±2.4**	81.3 ±3.3	72.8 ±3.4	**90.0 ±3.0**	79.4 ±2.5	67.8	-
ReasonMed-7B (1epoch)	65.3 ±1.3	62.3 ±0.8	82.0 ±5.5	74.8 ±3.7	80.0 ±2.5	81.3 ±3.3	**74.0 ±3.4**	86.0 ±3.5	79.0 ±2.5	67.7	-
ResponseMed-7B	67.5 ±1.3	60.9 ±0.8	80.0 ±5.7	74.8 ±3.7	77.4 ±2.6	84.0 ±3.1	71.1 ±3.5	88.0 ±3.3	76.5 ±2.6	67.0	225.2
CoTMed-7B	66.3 ±1.3	64.7 ±0.7	80.0 ±5.7	75.6 ±3.7	79.6 ±2.5	82.1 ±3.2	71.7 ±3.4	86.0 ±3.5	79.9 ±2.6	69.1	555.4
ReasonMed-7B	66.9 ±1.3	**65.1 ±0.7**	82.0 ±5.5	75.6 ±3.7	79.3 ±2.5	79.2 ±3.4	73.4 ±3.4	85.0 ±3.6	**80.9 ±2.4**	**69.6**	626.0

Table 4: Performance Comparison of Various Models on MedQA, MedMCQA, PubMedQA, and MMLU Benchmarks with Total Accuracy and Average Token Length.

also compared model performances trained for different epochs:

One Epoch Training: Models trained for one epoch showed promising yet suboptimal performance compared to their three-epoch counterparts. CoTMed-1epoch achieved an overall accuracy of 67.8%, slightly outperforming ReasonMed-7B-1epoch (67.7%) and significantly surpassing ResponseMed-7B-1epoch (64.8%).

Three Epoch Training: Models trained for three epochs consistently improved across benchmarks, clearly illustrating the benefit of extended training. The enhancements , whose overall accuracy improved from 67.71% (1 epoch) to 69.63% (3 epochs).

Under limited training steps, the CoTMed-7B model outperforms ReasonMed-7B; however, as the number of training steps increases, ReasonMed-7B ultimately surpasses CoTMed-7B by 0.54%. Additional training may enable the model to more effectively learn the internal connections between complex chain-of-thought reasoning and concise summarization, resulting in further performance gains.

Analysis of Average Token Length To obtain these averages, we ran each model in inference mode on all test set questions and computed the mean number of output tokens. CoTMed-7B (555 tokens) and ReasonMed-7B (626 tokens) generate substantially more content than ResponseMed-7B (225 tokens), reflecting deeper reasoning at the cost of verbosity. Compared to HuatuoGPT-o1-7B (446 tokens), our CoTMed and ReasonMed models exhibit even more extensive thought processes. Although ResponseMed-7B produces fewer tokens, it still outperforms the HuatuoGPT-o1 models in overall accuracy, highlighting the importance of dataset size and quality in model performance.

Compared to other biomedical LLMs such as BioMistral-7B, Llama3-OpenBioLLM-8B, and HuatuoGPT-o1, our ReasonMed-7B demonstrates outstanding medical QA performance, achieving the highest overall metrics. It outperforms the best same-size model by 4.17% and even surpasses certain ten-billion-parameter models on several benchmarks (see Appendix). These results underscore the importance of both dataset quality and scale, as well as the value of explicit multi-step reasoning in medical QA. Moreover, with additional training steps, the model is better able to internalize the relationship between detailed reasoning chains and concise response summaries, which significantly enhances its overall performance.

6 Conclusion

In this work, we introduced the ReasonMed, the largest open-source medical reasoning dataset, designed to enhance the performance of reasoning models in complex medical QA tasks. Using a multi-agent framework, we generated, verified, and optimized 1.291 million reasoning paths, refining them into 370k high-quality examples. Through rigorous fine-tuning experiments, we demonstrated that incorporating explicit multi-step reasoning significantly improves model performance, with our hybrid approach combining Chain-of-Thought reasoning and summarization achieving the best results. Outperformed existing models, including those with larger parameter sizes. These findings highlight the importance of reasoning in medical QA and provide a scalable framework for further research in knowledge-intensive domains.

Limitations

Due to constraints in computational resources, we did not extend our multiscale finetuning experiments to models larger than 7B parameters. While our hybrid ReasonMed-7B model outperforms many samesize and even some larger models on key benchmarks, it remains unclear how our dataset and finetuning strategies would scale when applied to stateoftheart models in the 10B-100B parameter range. Our data filtering (Verifier and Quality Ranker) and final quality assessment (Score Evaluator) rely exclusively on other large language models (Qwen-2.5-72B and GPT-4o). While these models are among the most advanced open-source, they may still harbor biases or systematic errors, which can occasionally result in misjudgments.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *Preprint*, arXiv:2402.00157.

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

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhishek Dutta and Yen-Che Hsiao. 2024. Adaptive reasoning and acting in medical language agents. *arXiv preprint arXiv:2410.10020*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

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)*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6.

Dong Huang, Jie M. Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. 2024. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *Preprint*, arXiv:2312.13010.

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*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Jack Krolik, Herprit Mahal, Feroz Ahmad, Gaurav Trivedi, and Bahador Saket. 2024. Towards leveraging large language models for automated medical qa evaluation. *Preprint*, arXiv:2409.01941.

Xiang Li, Duyi Pan, Hongru Xiao, Jiale Han, Jing Tang, Jiabao Ma, Wei Wang, and Bo Cheng. 2025. Dialogueagents: A hybrid agent-based speech synthesis framework for multi-party dialogue. *Preprint*, arXiv:2504.14482.

Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. Logical reasoning in large language models: A survey. *Preprint*, arXiv:2502.09100.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions? *Preprint*, arXiv:2207.08143.

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

OpenAI, :, Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, and 7 others. 2025. Competitive programming with large reasoning models. *Preprint*, arXiv:2502.06807.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference*

on Health, Inference, and Learning, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

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

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 replication journey: A strategic progress report – part 1. *Preprint*, arXiv:2410.18982.

Yi-Da Tang, Er-Dan Dong, and Wen Gao. 2024. Llms in medicine: The need for advanced evaluation systems for disruptive technologies.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

T. Zhao, S. Wang, C. Ouyang, M. Chen, C. Liu, J. Zhang, L. Yu, F. Wang, Y. Xie, J. Li, F. Wang, S. Grunwald, B. M. Wong, F. Zhang, Z. Qian, Y. Xu, C. Yu, W. Han, T. Sun, and 32 others. 2024. Artificial intelligence for geoscience: Progress, challenges, and perspectives. *Innovation (Camb)*, 5(5):100691.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pages 195–204. PMLR.

A Appendix

A.1 Ethical Statement

The ReasonMed-7B model presented in this paper has demonstrated strong performance in handling complex medical reasoning tasks. Nonetheless, it still carries a risk of generating inaccurate information, incomplete explanations, or hallucinations, which could potentially mislead users. Therefore, we strongly advise against the direct use of this model in clinical settings or any real-world applications where errors might lead to significant negative consequences. To ensure responsible usage, we restrict the model exclusively to academic research purposes. It is essential for users to recognize and respect these guidelines, thus avoiding situations in which the dissemination of incorrect medical information could compromise patient safety, treatment accuracy, or clinical judgment.

article [most]tcolorbox minted

A.2 Component Prompt Design

CoT Generate This component is used to generate medical MCQ analysis prompts with detailed chain thinking (CoT) to guide the model for step-by-step reasoning.

CoT Generate

"""

You are a highly knowledgeable medical expert. You are provided with a clinical multiple-choice question along with several candidate answers. Your task is to carefully analyze the clinical scenario and each option by following these steps:

1. Restate the question in your own words.
2. Highlight the key clinical details and relevant background information (e.g., pathophysiology, anatomy, typical presentations, diagnostic tests).
3. Evaluate each candidate answer, discussing supporting evidence and potential pitfalls.
4. Systematically rule out options that do not align with the clinical context.
5. Compare any remaining choices based on their merits.
6. Conclude with your final answer accompanied by a clear and concise summary of your reasoning.

Please note: Your response should be based solely on the current question and candidate answers. Do not consider any previous context or prior interactions.

Question:
{question}
Candidate Answers:
{options}
Please provide your detailed chain-of-thought reasoning followed by your final answer.
"""

Verifier This component is used to evaluate the chain-of-thoughts generated by the Multi-Agent system to determine whether their reasoning is correct and output JSON results.

Verifier

"""
You are a medical evaluation expert. Analyze if the Chain-of-Thought (CoT) analysis correctly leads to the answer.
[Question]
{question}
[Options]
{options$_str$}
[Correct Answer]
{answer}
[CoT Analysis]
{cot$_content$}
Evaluate the CoT analysis following these criteria:
1. Does the analysis correctly identify key clinical factors?
2. Are all options appropriately considered and evaluated?
3. Does the reasoning logically lead to the correct answer?
4. Are there any factual errors in medical knowledge?
Output a JSON object with:
- "verdict": "Correct" if the CoT analysis is valid and reaches the correct answer, otherwise "Error"
- "reason": Brief explanation of your evaluation (1-2 sentences)
"""

Response Summarizer This component is used to refine long-form CoT reasoning into concise summaries.

Response Summarizer

"""
Summarize the following chain-of-thought reasoning:
{cot}
"""

Quality Ranker This component is used to compare multiple CoT outputs, select the best two and give a brief justification for the other CoT Paths.

Quality Ranker

"""
You are a medical reasoning evaluator. Given the question, options, and known answer, review the following chains-of-thought (CoTs) labeled by their keys.
Select the two most sound and useful CoTs, then provide brief justifications for why each of the other CoTs were not chosen.
[Question]
{question}
[Options]
A) {optA}
B) {optB}
C) {optC}
D) {optD}
[Correct Answer]
{answer}
[CoTs]
{cot_block}
Respond with a JSON object with exactly two keys:
"top2": ["modelX_COTY", "modelZ_COTW"],
"reasons": <label>: <one-sentence justification> for every CoT not in top2
"""

Error Refiner This component refines the top two selected reasoning paths by incorporating targeted corrections for the error-prone points highlighted by the Verifier.

Error Refiner

"""
You are an expert clinician-educator AI tutor. Your mission is to generate an exceptionally comprehensive, in-depth chain-of-thought explanation that rigorously justifies the correct answer for the given clinical MCQ, while specifically addressing and integrating provided error feedback to eliminate previous reasoning flaws. Adhere closely to these instructions to maximize completeness:
1. **Error-Driven Refinement**
- Review the provided **Error Reasons from Other Attempts**.
- Identify logical gaps, factual mistakes, omissions, or misleading inferences in the

original chainofthought.
- Explicitly incorporate corrections and clarifications derived from these error reasons.
2. **Structured, Layered Reasoning**
Organize your explanation into clear sections:
a. Restate the question in your own words.
b. Highlight the key clinical details and relevant background information (e.g., pathophysiology, anatomy, typical presentations, diagnostic tests).
c. Evaluate each candidate answer, discussing supporting evidence and potential pitfalls.
d. Systematically rule out options that do not align with the clinical context.
e. Compare any remaining choices based on their merits.
f. Conclude with your final answer accompanied by a clear and concise summary of your reasoning.
Inputs
- **Question:** '{question}'
- **Options:** '{options}'
- **Correct Answer:** '{answer}'
- **Original Chain-of-Thought:** '{original_cot}'
- **Error Reasons from Other Attempts:** '{error_reasons}'

Output:
Please optimized Original Chain-of-Thought. Ensure that you explicitly address and rectify each error reason provided.
"""

Score Evaluator This component evaluates the quality of a models answer to each question and assigns a corresponding score.

Score Evaluator

"""
You are a medical reasoning evaluator. Assess the following response based on the following criteria:
1. **Clinical accuracy**: Does the response correctly incorporate medical facts, clinical guidelines, and evidence-based practices? Are the clinical details provided accurate, relevant, and appropriate for the given situation?
2. **Logical reasoning**: Does the response logically follow the reasoning process required to arrive at the answer? Is the reasoning chain coherent and well-supported by evidence or clinical knowledge?
3. **Factual correctness**: Are there any factual errors in the response? Are all statements factually correct and consistent with established medical knowledge?
4. **Completeness**: Does the response cover all necessary aspects of the question? Is it thorough and detailed, addressing the key points without missing critical information?
[Question]
{question}
[Response]
{response}
Please evaluate the response on the above criteria and provide a JSON object with two keys:
"score": integer between 1 and 10,
"justification": A concise explanation of your score.
"""

A.3 Additional Experiments

In Table 5, we presented pairwise (1-vs-1) differences among DeepSeek-R1-Distill-Llama-70B, HuatuoGPT-o1-70B, and Qwen2.5-72B, showing for each pair the count of questions one model answered correctly but the other did not. To further explore complementary coverage, Table 6 summarizes the one-vs-two scenario: for each model, the number of questions it missed while the other two both answered correctly. DeepSeek-R1-Distill-Llama-70B failed only 3,430 (1.76%) questions that HuatuoGPT-o1-70B and Qwen2.5-72B both got right; HuatuoGPT-o1-70B missed 9,352 (4.80%); and Qwen2.5-72B missed 5,280 (2.71%), out of 194,925 total. Together, these results confirm that each model contributes unique strengths and gaps, underscoring the value of ensemble or multi-agent approaches in medical QA.

Table 7 compares the performance of various LLaMA3.1 and Qwen2.5 models on several key medical benchmarks, including MedQA, MedMCQA, PubMedQA, and six sub-domains of MMLU.Our model,ReasonMed-7B, consistently

Comparison	Correct by Model 1 but Incorrect by Model 2	Incorrect by Model 1 but Correct by Model 2	Total Questions
DeepSeek-R1-Distill-Llama-70B vs HuatuoGPT-o1-70B	8,168 (4.19%)	27,339 (14.03%)	194,925
DeepSeek-R1-Distill-Llama-70B vs Qwen2.5-72B	19,017 (9.76%)	23,267 (11.94%)	194,925
Qwen2.5-72B vs HuatuoGPT-o1-70B	10,018 (5.14%)	24,939 (12.79%)	194,925

Table 5: Pairwise (1-vs-1) Knowledge Domain Differences among the three models.

Model	Questions Missed by This Model but Correct by Both Others	Total Questions
DeepSeek-R1-Distill-Llama-70B	3,430 (1.76%)	194,925
HuatuoGPT-o1-70B	9,352 (4.80%)	194,925
Qwen2.5-72B	5,280 (2.71%)	194,925

Table 6: Collective (1-vs-2) Miss Rates: questions each model failed while the other two both answered correctly.

	MedQA	MedMCQA (val)	PubMedQA	MMLU						Total Acc
				Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	
Dataset_Count	1273	4183	1000	135	265	144	173	100	272	-
LLaMA3.1-70B	76.8 ±0.1	67.9 ±0.7	77.4 ±0.2	**81.5 ±0.3**	**89.1 ±0.2**	**96.5 ±0.1**	**80.9 ±0.3**	90.0 ±0.3	**93.0 ±0.2**	72.9
Qwen2.5-14B	75.6 ±0.1	63.4 ±0.8	**77.6 ±0.2**	75.6 ±0.4	84.9 ±0.2	88.9 ±0.3	75.7 ±0.3	90.0 ±0.3	84.2 ±0.2	69.0
Qwen2.5-32B	79.3 ±0.1	67.6 ±0.7	**77.6 ±0.2**	79.3 ±0.3	86.8 ±0.2	93.8 ±0.2	79.8 ±0.3	91.0 ±0.3	87.5 ±0.2	72.6
Qwen2.5-72B	**81.5 ±0.1**	**71.2 ±0.1**	76.4 ±0.2	75.6 ±0.4	86.8 ±0.2	93.8 ±0.2	77.5 ±0.3	**92.0 ±0.3**	88.2 ±0.2	**75.6**
QwQ-32B	78.1 ±0.1	65.5 ±0.7	76.4 ±0.2	75.6 ±0.4	86.8 ±0.2	93.8 ±0.2	77.5 ±0.3	**92.0 ±0.3**	88.2 ±0.2	72.0

Table 7: Performance Comparison of LLaMA3.1 and Qwen2.5 Series Models(over 10B) on MedQA, MedMCQA, PubMedQA, and MMLU Benchmarks.

achieves competitive performance and even outperforms much larger models on several benchmarks. Most notably, on PubMedQA, ReasonMed-7B surpasses LLaMA3.1-70B by 4.60%, demonstrating the effectiveness of our dataset and fine-tuning strategy. Moreover, the model delivers strong results on MedMCQA and multiple MMLU medical subdomains, highlighting its superior reasoning capabilities in knowledge-intensive tasks. These findings validate our multi-stage data refinement and Chain-of-Thought optimization strategy, establishing ReasonMed-7B as a new state-of-the-art among sub-10B models, and narrowing the performance gap with frontier 70B+ models.

A.4 Example of model output

Prompt

Please answer the following multiple-choice question:
Which of the following is not true about glomerular capillaries?
A. The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it
B. Glucose concentration in the capillaries is the same as that in glomerular filtrate
C. Constriction of afferent arteriole decreases the blood flow to the glomerulus
D. Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it

ReasonMed (temperature 0.4)

To answer this question, let's break down the components step-by-step:

1. Restate the question: We need to identify which statement about glomerular capillaries is incorrect.

2. Key clinical details and relevant background information:
- Glomerular capillaries are part of the kidney's filtration system.
- They filter blood to form urine.
- Oncotic pressure is the pressure exerted by proteins in the blood.
- Glomerular filtrate is the fluid that passes through the glomerulus.
- Afferent arterioles supply blood to the glomerulus.
- Hematocrit refers to the proportion of blood volume occupied by red blood cells.

3. Evaluate each candidate answer:
- "The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it": This is true because as fluid

leaves the capillaries, it carries away proteins, reducing the oncotic pressure.
- "Glucose concentration in the capillaries is the same as that in glomerular filtrate": This is false. Glucose is actively reabsorbed in the proximal tubule, so its concentration in the filtrate is lower than in the capillaries.
- "Constriction of afferent aeriole decreases the blood flow to the glomerulas": This is true. Constriction reduces blood flow, which can decrease filtration rate.
- "Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it": This is true. As fluid leaves the capillaries, it carries away proteins, including red blood cells, reducing hematocrit.

4. Rule out options that do not align with the clinical context:
- We can rule out the first, third, and fourth options as they align with known physiological processes.

5. Compare any remaining choices:
- The second option stands out as the only one that contradicts known physiology.

6. Final answer:
The statement that is not true about glomerular capillaries is: "Glucose concentration in the capillaries is the same as that in glomerular filtrate." This is because glucose is actively reabsorbed in the proximal tubule, leading to a lower concentration in the filtrate compared to the capillaries.
</think>
The chain-of-thought reasoning breaks down the process of identifying the incorrect statement about glomerular capillaries step-by-step. It involves restating the question, providing key clinical details and background information, evaluating each candidate answer, ruling out options that do not align with the context, comparing remaining choices, and arriving at the final answer. The incorrect statement identified is that the glucose concentration in the capillaries is the same as that in glomerular filtrate, which is false because glucose is actively reabsorbed in the proximal tubule.

ReasonMed (temperature 0.5)

To answer this question, let's break down the key concepts related to glomerular capillaries and evaluate each statement accordingly.

1. **Restating the Question**: We need to identify which statement about glomerular capillaries is incorrect.

2. **Key Clinical Details and Background Information**:
- Glomerular capillaries are part of the kidney's filtration system.
- They are responsible for filtering blood to form urine.
- Key processes include filtration, reabsorption, and secretion.
- Oncotic pressure, glucose concentration, blood flow, and hematocrit are important parameters in this context.

3. **Evaluating Each Candidate Answer**:
- **The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it**: This statement is true. As blood passes through the glomerular capillaries, proteins are filtered out, reducing the oncotic pressure of the fluid leaving the capillaries compared to the fluid entering them.
- **Glucose concentration in the capillaries is the same as that in glomerular filtrate**: This statement is false. Glucose is actively reabsorbed in the proximal tubule, so the glucose concentration in the capillaries is higher than in the glomerular filtrate.
- **Constriction of afferent arteriole decreases the blood flow to the glomerulus**: This statement is true. Constriction of the afferent arteriole reduces blood flow into the glomerulus, thereby decreasing glomerular filtration rate (GFR).
- **Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it**: This statement is true. As blood passes through the glomerular capillaries, plasma proteins are filtered out, leading to a decrease in hematocrit in the fluid leaving the capillaries.

4. **Systematically Ruling Out Options**:
- The first, third, and fourth statements align with the known physiology of glomerular capillaries.
- The second statement contradicts the known physiology, as glucose is actively reabsorbed, making its concentration higher in the capillaries than in the glomerular filtrate.

5. **Comparing Remaining Choices**:
- Only the second statement does not align with the expected physiological behavior of glomerular capillaries.

6. **Final Answer**:
- The statement that is not true about glomerular capillaries is: "Glucose concentration in the capillaries is the same as that in glomerular filtrate."

Summary of Reasoning:
The correct answer is that the glucose concentration in the capillaries is not the same as that in the glomerular filtrate. This is because glucose is actively reabsorbed in the proximal tubule, resulting in a higher glucose concentration in the capillaries compared to the filtrate. All other statements accurately describe aspects of glomerular capillary function.

806

ReasonMed (temperature 0.6)

To determine which statement is not true about glomerular capillaries, let's analyze each option step-by-step.

1. **The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it:**
- Oncotic pressure is the colloid osmotic pressure exerted by proteins in the blood plasma. In the glomerulus, as fluid passes through the fenestrated endothelium and the glomerular basement membrane, some plasma proteins are too large to pass through. Therefore, the oncotic pressure of the fluid that leaves the capillaries (glomerular filtrate) is indeed lower than the oncotic pressure of the fluid entering the capillaries (plasma). This statement is true.

2. **Glucose concentration in the capillaries is the same as that in glomerular filtrate:**
- Glucose is freely filtered by the glomerulus, meaning that it passes from the capillaries into the glomerular filtrate without being reabsorbed or secreted. Thus, the glucose concentration in the capillaries should be equal to the glucose concentration in the glomerular filtrate. This statement is true.

3. **Constriction of afferent arteriole decreases the blood flow to the glomerulus:**
- The afferent arteriole supplies blood to the glomerulus. Constriction of this arteriole reduces the amount of blood flowing into the glomerulus, thereby decreasing the blood flow. This statement is true.

4. **Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it:**
- Hematocrit refers to the proportion of red blood cells in a given volume of blood. As fluid leaves the capillaries and enters the glomerular filtrate, plasma proteins are retained in the capillaries while water and small molecules are filtered out. Since plasma proteins contribute to the hematocrit, the hematocrit of the fluid leaving the capillaries (glomerular filtrate) is actually higher than the hematocrit of the fluid entering the capillaries (plasma). This statement is false.

Based on this analysis, the statement that is not true about glomerular capillaries is:

Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it.
</think>
The reasoning process involves evaluating statements about glomerular capillaries to identify which one is not true.

807

808

1. The oncotic pressure of the fluid leaving the capillaries is lower than that of the fluid entering due to the retention of plasma proteins. This statement is true.
2. Glucose concentration in the capillaries is the same as in glomerular filtrate because glucose is freely filtered. This statement is true.
3. Constriction of the afferent arteriole decreases blood flow to the glomerulus, which is accurate. This statement is true.
4. The hematocrit of the fluid leaving the capillaries is actually higher than that of the fluid entering because plasma proteins remain in the capillaries. This statement is false.

Thus, the incorrect statement is that the hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it.
<system />

809

CoTMed (temperature 0.4)

Let's break down the question and evaluate each candidate answer step-by-step.

Restated Question:
Which statement about glomerular capillaries is incorrect?

Key Clinical Details and Relevant Background Information:
1. **Glomerular Filtration**: Glomerular capillaries are part of the kidney's filtration system where blood is filtered to form urine.
2. **Oncotic Pressure**: This is the osmotic pressure exerted by proteins in the blood plasma.
3. **Glucose Concentration**: Glucose is filtered through the glomerulus but is reabsorbed in the proximal tubule.
4. **Afferent and Efferent Arterioles**: These regulate blood flow into and out of the glomerulus.
5. **Hematocrit**: The proportion of blood volume occupied by red blood cells.

Evaluation of Each Candidate Answer:

810

1. **The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it.**
- **Supporting Evidence**: As blood passes through the glomerular capillaries, water and small molecules are filtered out, leaving behind a higher concentration of proteins (and thus higher oncotic pressure). Therefore, the oncotic pressure of the fluid leaving the capillaries is indeed higher than that of the fluid entering it.
- **Potential Pitfall**: This statement is incorrect because the oncotic pressure increases as the fluid leaves the capillaries.

2. **Glucose concentration in the capillaries is the same as that in glomerular filtrate.**
- **Supporting Evidence**: Glucose is freely filtered through the glomerulus, so the initial concentration in the capillaries should match the concentration in the glomerular filtrate.
- **Potential Pitfall**: While this is generally true, glucose is actively reabsorbed in the proximal tubule, so the concentration in the filtrate will be lower than in the capillaries after filtration.

3. **Constriction of afferent arteriole decreases the blood flow to the glomerulus.**
- **Supporting Evidence**: Constriction of the afferent arteriole reduces blood flow into the glomerulus, thereby decreasing glomerular filtration rate (GFR).
- **Potential Pitfall**: This statement is correct and aligns with physiological principles.

4. **Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it.**
- **Supporting Evidence**: As blood passes through the glomerular capillaries, water and small molecules are filtered out, leaving behind a higher concentration of red blood cells (and thus a lower hematocrit). Therefore, the hematocrit of the fluid leaving the capillaries is indeed lower than that of the fluid entering it.

811

- **Potential Pitfall**: This statement is correct and aligns with physiological principles.

Systematic Ruling Out Options:
- Option 1 is incorrect because the oncotic pressure increases as the fluid leaves the capillaries.
- Option 2 is generally true but has a potential pitfall regarding active reabsorption.
- Option 3 is correct.
- Option 4 is correct.

Comparison of Remaining Choices:
- Option 1 is clearly incorrect based on the physiological principles of oncotic pressure.

Final Answer:
The statement that is not true about glomerular capillaries is:
- **The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it.**

This is because the oncotic pressure increases as the fluid leaves the capillaries due to the filtration of water and small molecules.

CoTMed (temperature 0.5)

Let's break down the question and evaluate each candidate answer step-by-step:

1. **Restate the Question:**
- Which statement about glomerular capillaries is incorrect?

2. **Key Clinical Details and Background Information:**
- Glomerular capillaries are part of the kidney's filtration system.
- They are responsible for filtering blood to form urine.
- Key processes include filtration, reabsorption, and secretion.
- Oncotic pressure, glucose concentration, blood flow, and hematocrit are important parameters.

3. **Evaluate Each Candidate Answer:**

- **The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it:**
- Oncotic pressure is primarily due to proteins in the plasma.
- As fluid leaves the capillaries into the Bowman's capsule, most of the proteins remain in the capillaries.
- Therefore, the oncotic pressure of the fluid leaving the capillaries is indeed lower than that of the fluid entering it.
- This statement is true.

- **Glucose concentration in the capillaries is the same as that in glomerular filtrate:**
- Glucose is filtered from the blood into the glomerular filtrate.
- However, glucose is also reabsorbed in the proximal tubule.
- Initially, the glucose concentration in the capillaries and filtrate might be similar, but as filtration occurs, the filtrate will have more glucose.
- This statement is false because the glucose concentration in the capillaries is typically higher than in the glomerular filtrate due to ongoing reabsorption.

- **Constriction of afferent arteriole decreases the blood flow to the glomerulus:**
- The afferent arteriole supplies blood to the glomerulus.
- Constriction of this arteriole reduces blood flow to the glomerulus.
- This statement is true.

- **Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it:**
- Hematocrit refers to the proportion of red blood cells in the blood.
- As fluid leaves the capillaries, the remaining fluid has a higher concentration of proteins and a lower concentration of red blood cells.
- Therefore, the hematocrit of the fluid leaving the capillaries is indeed lower than that of the fluid entering it. - This statement is true.

4. **Systematically Rule Out Options:**
- We can rule out the statements about oncotic pressure, constriction of the afferent arteriole, and hematocrit as they are all true.

5. **Compare Remaining Choices:**
- The only remaining choice is the statement about glucose concentration.

6. **Final Answer:**
- The statement "Glucose concentration in the capillaries is the same as that in glomerular filtrate" is not true.

Conclusion:
The correct answer is: **Glucose concentration in the capillaries is the same as that in glomerular filtrate.** This statement is false because the glucose concentration in the capillaries is typically higher than in the glomerular filtrate due to ongoing reabsorption in the proximal tubule.

CoTMed (temperature 0.6)

To answer this question, let's break down the options step by step:

1. **Restating the Question**: We need to identify which statement about glomerular capillaries is incorrect.

2. **Key Clinical Details and Background Information**:
- Glomerular capillaries are part of the kidney's filtration system.
- They are involved in the formation of urine through the process of glomerular filtration.
- The glomerular capillaries have a unique structure that allows for selective filtration of blood components.
- Oncotic pressure, glucose concentration, and hematocrit are important factors in understanding the function of glomerular capillaries.

3. **Evaluating Each Candidate Answer**:
- **Option 1: The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it**
- This statement is true. Oncotic pressure is the osmotic pressure exerted by proteins in the blood. As fluid passes through the glomerular capillaries, some of the plasma proteins are too large to pass through the filtration barrier, so they remain in the capillaries. Therefore, the oncotic pressure of the fluid leaving the capillaries is indeed lower than that of the fluid entering it.

- **Option 2: Glucose concentration in the capillaries is the same as that in glomerular filtrate**
- This statement is false. Glucose is filtered from the blood into the glomerular filtrate, but the filtration process is selective. The glomerular capillaries reabsorb glucose back into the bloodstream, so the glucose concentration in the capillaries is higher than in the glomerular filtrate.

- **Option 3: Constriction of afferent arteriole decreases the blood flow to the glomerulus**
- This statement is true. The afferent arteriole supplies blood to the glomerulus. Constriction of this arteriole reduces blood flow to the glomerulus, thereby decreasing the glomerular filtration rate (GFR).

- **Option 4: Hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it**
- This statement is true. Hematocrit refers to the proportion of blood volume occupied by red blood cells. As fluid passes through the glomerular capillaries, plasma proteins are retained, while water and small molecules are filtered out. This results in a lower hematocrit in the fluid leaving the capillaries compared to the fluid entering them.

4. **Ruling Out Options**:
- Options 1, 3, and 4 are all true statements about glomerular capillaries.
- Option 2 is the only false statement among the given options.

5. **Comparing Remaining Choices**:

- Since we have identified that Option 2 is the only false statement, there is no need for further comparison.

6. **Final Answer**:
- The correct answer is **Option 2: Glucose concentration in the capillaries is the same as that in glomerular filtrate**. This statement is false because the glomerular capillaries reabsorb glucose, resulting in a higher glucose concentration in the capillaries compared to the glomerular filtrate.

818

ResponseMed (temperature 0.4)

The question asks which statement about glomerular capillaries is incorrect. Key details include that glomerular capillaries are part of the kidney's filtration system, where blood is filtered to produce urine, and that oncotic pressure affects fluid movement across capillary walls.

Evaluating the candidate answers:
1. The statement about oncotic pressure is incorrect because the oncotic pressure of fluid leaving the capillaries is actually greater than that of the fluid entering, due to the reabsorption of proteins.
2. The statement regarding glucose concentration is true; glucose is filtered and then reabsorbed, leaving its concentration unchanged.
3. The statement about afferent arteriole constriction decreasing blood flow to the glomerulus is also true.
4. The statement about hematocrit is true as well; the hematocrit of the fluid leaving the capillaries is lower due to plasma filtration.

Thus, the incorrect statement is that "The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it." This is because the oncotic pressure increases as proteins are retained in the capillaries.

Final answer: The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it. This statement is incorrect as the oncotic pressure actually increases due to protein retention.

820

ResponseMed (temperature 0.5)

The incorrect statement regarding glomerular capillaries is that "The oncotic pressure of the fluid leaving the capillaries is less than that of fluid entering it." This is false because oncotic pressure increases as fluid moves through the glomerulus due to the removal of plasma proteins, leading to a higher oncotic pressure in the capillary blood compared to the glomerular filtrate.

The other statements are accurate:
1. Glucose concentration in capillaries is the same as in glomerular filtrate initially, but it is fully reabsorbed, making this statement somewhat misleading.
2. Constriction of the afferent arteriole does decrease blood flow to the glomerulus, which is correct.
3. The hematocrit of the fluid leaving the capillaries is indeed lower than that entering, as plasma is filtered out, making this statement true.

In summary, the first statement is the only one that is definitively incorrect based on the principles of glomerular filtration. Understanding these dynamics is crucial for interpreting renal function and managing related medical conditions.

821

ResponseMed (temperature 0.6)

To determine which statement about glomerular capillaries is incorrect, we analyzed several options based on the physiology of the glomerulus.

1. **Oncotic Pressure**: The oncotic pressure of fluid leaving the capillaries is indeed lower than that entering, as some proteins are retained, making this statement true.
2. **Glucose Concentration**: Glucose is freely filtered, so its concentration is the same in both capillary blood and glomerular

819

822

filtrate, confirming this statement is true.
3. **Afferent Arteriole Constriction**: Constriction of the afferent arteriole reduces blood flow to the glomerulus, which is also true.
4. **Hematocrit**: The hematocrit of the fluid leaving the capillaries is higher than that entering because plasma is filtered out, making this statement false.

Thus, the incorrect statement is that the hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it.

Final Answer: The incorrect statement is that the hematocrit of the fluid leaving the capillaries is less than that of the fluid entering it. $\boxed{D}$