Med-PRM: Medical Reasoning Models with Stepwise, Guideline-verified Process Rewards

Jaehoon Yun[1,4,5*], **Jiwoong Sohn**[1,2*], **Jungwoo Park**[1,4*], **Hyunjae Kim**[3], **Xiangru Tang**[3], **Yanjun Shao**[3], **Yonghoe Koo**[6], **Minhyeok Ko**[5], **Qingyu Chen**[3], **Mark Gerstein**[3], **Michael Moor**[2†], **Jaewoo Kang**[1,4†]
[1] Korea University, [2] ETH Zürich, [3] Yale University, [4] AIGEN Sciences,
[5] Hanyang University College of Medicine, [6] University of Ulsan College of Medicine

Abstract

Large language models have shown promise in clinical decision making, but current approaches struggle to localize and correct errors at specific steps of the reasoning process. This limitation is critical in medicine, where identifying and addressing reasoning errors is essential for accurate diagnosis and effective patient care. We introduce Med-PRM, a process reward modeling framework that leverages retrieval-augmented generation to verify each reasoning step against established medical knowledge bases. By verifying intermediate reasoning steps with evidence retrieved from clinical guidelines and literature, our model can precisely assess the reasoning quality in a fine-grained manner. Evaluations on five medical QA benchmarks and two open-ended diagnostic tasks demonstrate that Med-PRM achieves state-of-the-art performance, with improving the performance of base models by up to 13.50% using Med-PRM. Moreover, we demonstrate the generality of Med-PRM by integrating it in a plug-and-play fashion with strong policy models such as Meerkat, achieving over 80% accuracy on MedQA for the first time using small-scale models of 8 billion parameters. Our code and data are available at Med-PRM.github.io.

1 Introduction

Clinical decision making (CDM) is a complex, multi-step process involving the assessment of patient symptoms, retrieval of relevant clinical evidence, and formulation of diagnostic and treatment strategies. Unlike simple factual recall, CDM requires integrating diverse clinical findings and dynamically refining hypotheses as new information becomes available. Effective CDM entails not only selecting the most probable differential diagnoses but also determining what additional information is needed to reduce uncertainty and guide the next best diagnostic and therapeutic steps in a patient's clinical trajectory (Moor et al., 2023a).

While CDM spans a broad sequence of clinical decisions, a core subcomponent is the step-by-step reasoning that underlies each decision point. This sequential structure makes CDM reasoning well-suited to process reward modeling (PRM) (Lightman et al., 2023; Uesato et al., 2022; Setlur et al., 2024), which evaluates and rewards intermediate steps of a process rather than solely its final outcome. In medical practice, sound intermediate reasoning is critical to ensuring safety, reliability, and adherence to the standard of care. This creates a strong clinical motivation for models that support stepwise verification and feedback.

Recent advances in large language models (LLMs) have substantially improved performance in medical applications through pre-training (Chen et al., 2023; Moor et al., 2023b), post-training (Kim et al., 2025), retrieval (Zakka et al., 2024; Jeong et al., 2024; Sohn et al., 2025), tool augmentation, and agentic systems (Tang et al., 2024; Schmidgall et al., 2024). More recently, emerging reasoning models (OpenAI, 2024; DeepSeek-AI et al., 2025) have demonstrated the ability to decompose complex tasks into interpretable steps and exhibit metacognitive skills such as planning and error correction. However, such abilities are underexplored in clinical domains, where transparency, robustness, and alignment with medical standards are critical for delivering high-quality care.

Despite PRM's potential for medical reasoning, its application to the medical domain poses key challenges. Chief among these is the need for high-quality, step-level supervision, which is both expensive and labor-intensive to obtain. Early studies (Lightman et al., 2023) relied on human annotation, which is not scalable (Setlur et al., 2024). More recent works employed automatic labeling strategies such as Monte Carlo Tree Search

[*] Equal contribution.
[†] Corresponding authors.

(MCTS) (Wang et al., 2024b), which estimate the quality of reasoning step quality based on the probability of reaching the correct final answer from that step. Notably, MedS^3 (Jiang et al., 2025), a domain-specific PRM for clinical QA, also adopts an MCTS-based approach. However, these strategies often undervalue early reasoning steps that are logically sound but fail to lead to the correct outcome. This limitation is especially problematic as penalizing valid early steps can distort the learning signal and ultimately hinder the model's ability to evaluate intermediate reasoning accurately.

Second, medical reasoning requires extensive domain knowledge that may not be fully captured within language model's parameters alone. Thus, it necessitates a robust method to incorporate medical knowledge to generate factual, evidence-based outcomes and prevent hallucinations. In particular, training reward models solely on labels without medical context is insufficient for learning the rationale behind those labels. To overcome this, it is essential to provide relevant medical information, such as clinical guidelines, during training, enabling a more accurate interpretation of stepwise reward signals grounded in medical reasoning.

To address these challenges, we propose Med-PRM, a retrieval-augmented process reward modeling framework for clinical reasoning. Our method employs a RAG-AS-A-JUDGE approach to perform stepwise evaluation conditioned on both the clinical question and retrieved medical documents. This retrieval-augmented evaluation aligns more closely with expert physician annotations than sampling-based auto-labeling methods used during training. By incorporating relevant clinical knowledge at both the training and inference stages, Med-PRM enables more accurate assessment of intermediate reasoning and outperforms existing PRM baselines by an average of 3.44% across seven medical benchmarks.

Our experiments demonstrate that test-time scaling, where Med-PRM is used as a verifier alongside a fine-tuned policy model, achieves state-of-the-art performance. Notably, Med-PRM exhibits strong plug-and-play generality: when applied to top-performing models such as UltraMedical (Zhang et al., 2024b), which is trained on data costing approximately $20,000, our reward model, trained on a curated dataset costing less than $20, further achieves superior performance, highlighting the cost-efficiency and scalability of our approach.

Our contributions are as follows:

1. We propose Med-PRM, a retrieval-augmented process reward modeling framework that evaluates each reasoning step in the context of both the clinical question and retrieved evidence, enabling fine-grained and evidence-grounded assessment.

2. We demonstrate that Med-PRM achieves state-of-the-art performance across six out of seven medical QA benchmarks, outperforming all baseline language, reasoning, and medical models. Our verifier improves base model performance by up to 13.50% at test time and reaches 80.35% on MedQA (4 options) using only 8B-parameter models.

3. Through in-depth qualitative analysis and collaboration with medical experts, we show that Med-PRM closely aligns with clinical experts, addressing key limitations of prior training methods of PRMs in both logical consistency and factual accuracy.

2 Related Work

For a more detailed overview of related work, refer to Appendix E. LLMs have shown increasing proficiency in medical reasoning, effectively handling domain-specific terminology and multimodal data. Using corpora like PubMed, MIMIC-III/IV (Johnson et al., 2023), and UMLS, recent models go beyond surface-level recall to complex inference.

Med-PaLM (Singhal et al., 2023a) demonstrates promising performance on expert-level medical QA benchmarks, while methods such as CoT prompting (Wei et al., 2023; Xu et al., 2024; Kim et al., 2025), agentic frameworks (Kim et al., 2024; Tang et al., 2024; Schmidgall et al., 2024), and PRM (Jiang et al., 2025) further enhance reasoning. Reinforcement learning (Wang et al., 2024b) and verifier feedback (Chen et al., 2024) have been applied to refine reasoning traces, emphasizing the need for stepwise supervision.

MedS^3(Jiang et al., 2025) applies PRM using MCTS-based auto-labeling. Our approach, Med-PRM, also leverages process-level rewards but differs by incorporating retrieval-augmented generation and an LLM-as-a-Judge framework. As shown in Section 6 and Section 7, this yields superior performance compared to MCTS-based methods.

Hao et al. (2024) explore the use of LLMs as verifiers for CoT reasoning. Med-PRM adopts RAG-AS-A-JUDGE for stepwise supervision, diverging

from earlier PRM approaches that rely on automatic scoring. In mathematics, RAG-PRM (Zhu et al., 2025) has been proposed to retrieve similar QA pairs for few-shot prompting with PRM. In contrast, Med-PRM retrieves medical knowledge and evidence, enabling integration of diverse sources like textbooks or clinical guidelines.

3 Preliminaries

3.1 Reward Model

Reward models have emerged as central to advancing LLMs beyond pre-training and fine-tuning, driven by two key developments in reinforcement learning (RL) and test-time compute scaling. RL methods such as Proximal Policy Optimization (PPO) rely on reward functions to optimize model behavior in settings where ground-truth supervision is sparse or costly. Additionally, test-time strategies like best-of-N (Lightman et al., 2023) have proven effective as alternatives to majority voting, such as self-consistency (Wang et al., 2023), using reward models to rank and select high-quality outputs. These trends highlight the growing importance of accurate, context-aware reward models not only during training but also at inference time.

Outcome Reward Model (ORM) Given a question q and a model-generated reasoning trace S the ORM assigns a sigmoid score $r_S \in [0, 1]$ indicating the correctness of the entire trace. ORM is trained with the following cross-entropy loss:

$$\mathcal{L}_{\text{ORM}} = -\left(y_S \log r_S + (1 - y_S) \log(1 - r_S)\right),$$

where y_S is the gold label of the reasoning trace S, $y_S = 1$ if S is correct, and $y_S = 0$ otherwise.

Process Reward Model (PRM) Given a reasoning trace $S = (s_1, s_2, \cdots, s_K)$ where K is the number of reasoning steps, a PRM assigns score $r_{s_i} \in [0, 1]$ for each step s_i. Gold labels $y_{s_i} \in \{0, 1\}$ indicate whether each step is correct. To compute these scores, the model predicts logits for the special tokens + (correct) and - (incorrect) appended to each reasoning step. The confidence score r_{s_i} is defined as the softmax probability of the + token over the logits of both tokens. The model is trained to minimize the cross-entropy loss over all reasoning steps:

$$\mathcal{L}_{\text{PRM}} = -\sum_{i=1}^{K} \left(y_{s_i} \log r_{s_i} + (1 - y_{s_i}) \log(1 - r_{s_i})\right)$$

PRM takes as input the concatenation of the question q and the reasoning trace S, and produces stepwise confidence scores as follows: $(r_{s_1}, r_{s_2}, \cdots, r_{s_K})$ And the minimum step score is defined as the score of the solution S:

$$\text{RM}(q, S) = r_S, \text{ where } r_S = \min(r_{s_1}, r_{s_2}, \cdots, r_{s_K})$$

PRM Auto-Labeling Wang et al. (2024b) proposed an auto-labeling method to address the cost of human annotations y_{s_i}. Inspired by MCTS, a completer model generates N subsequent reasoning processes from each partial trace up to step s_i, producing sequences of the form $\{(s_{i+1,j}, \cdots, s_{K_j,j}, a_j)\}_{j=1}^{N}$, where a_j is the final answer of the j-th continuation. Let a^* denote the gold answer to the question q. A hard label is assigned to step s_i as follows:

$$y_{s_i}^{HE} = \begin{cases} 1 & \text{if } \exists j \text{ such that } a_j = a^*, \\ 0 & \text{otherwise.} \end{cases}$$

A soft label is computed as the empirical probability of reaching the correct answer:

$$y_{s_i}^{SE} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}(a_j = a^*)$$

This auto-labeling approach enables scalable PRM training without human supervision. However, it is prone to false negatives—particularly in complex questions—where factually correct and logically coherent intermediate steps are mislabeled as incorrect (or assigned low soft-label scores) simply because none of their sampled continuations lead to the correct final answer. To mitigate false negatives, we incorporate retrieval-based fact-checking into the labeling process, as described in Section 4.

4 Method

4.1 RAG-AS-A-JUDGE

We use RAG-AS-A-JUDGE with a labeling strategy that differs from conventional PRMs. Given a question q, golden answer a^*, a set of retrieved documents D, and a sequence of reasoning steps $S = (s_1, \ldots, s_K)$, RAG-AS-A-JUDGE performs binary classification on each step s_i to determine whether it is correct, producing labels $y_{s_i}^{RAG} \in \{0, 1\}$:

$$\text{RAG-AS-A-JUDGE}(D, q, a^*, S) = y_S^{RAG}$$

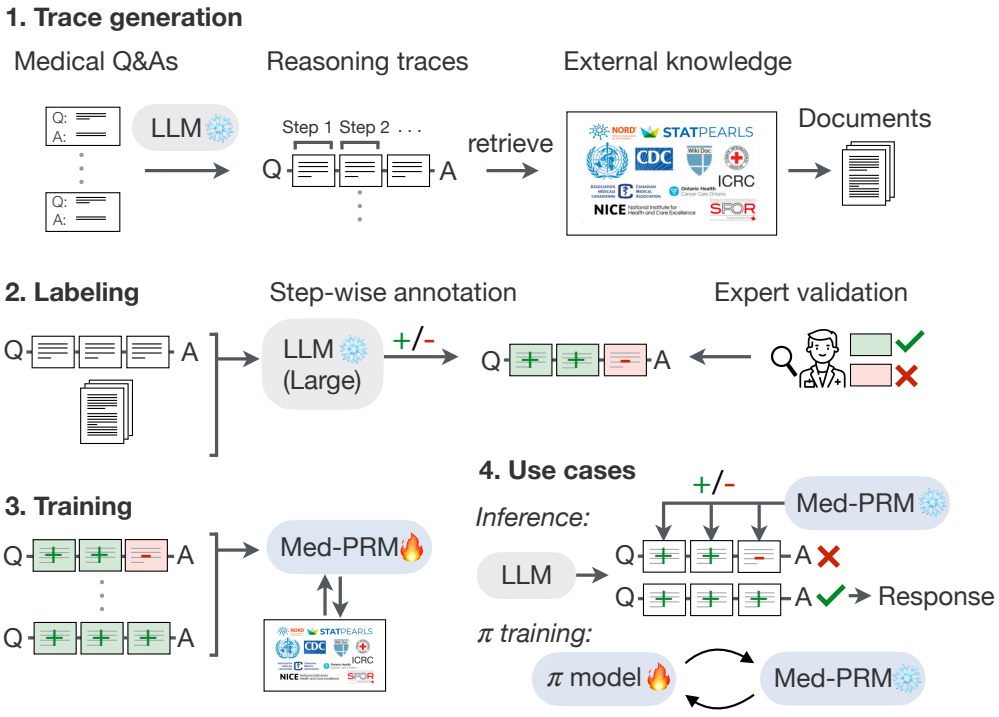


Figure 1: Overview of the MED-PRM. (1) An LLM generates reasoning traces for medical questions, and relevant documents are retrieved from external corpora. (2) A large LLM assigns stepwise labels (+/–) for each reasoning step. (3) These labeled traces are used to train the Med-PRM reward model. (4) Med-PRM reward model is used for inference-time evaluation or further train the policy model.

These labels y_S^{RAG} are plugged in $\mathcal{L}_{\text{PRM}}$ in order to train our PRM. Each score r_{s_i} reflects the likelihood that step s_i is correct, following Lightman et al. (2023). During both training and evaluation, PRM receives the same input: question q, retrieved documents D, and reasoning trace S. The key difference from prior PRM models is the inclusion of the retrieved documents D as part of the input.

$$\text{RM}(D, q, S) = r_S$$

To construct the retrieval query, we concatenate the question with the reasoning trace:

$$D = \text{Retriever}(q, S)$$

During training, the correct reasoning trace among multiple inferences is used as the query. During inference, a randomly sampled reasoning trace is used as the query instead. Further implementation details of document retrieval are provided in Appendix B.

Scaling Test-time Computation Following Lightman et al. (2023), we define a reasoning trace's final score as the minimum reward across its steps. The best-of-N approach selects an answer among the traces with the highest final score. Let $C = \{S_1, S_2, \cdots, S_N\}$ be the set of reasoning traces and $\{a_{S_1}, a_{S_2}, \cdots, a_{S_N}\}$ the corresponding answers. Then, we have

$$a_{rm} = a_{S_*}, \quad \text{where } S_* = \arg\max_{S_j \in C} r_{S_j},$$

and where r_{S_j} is a final score of reasoning trace S_j We also adopt a hybrid method following Li et al. (2023) that combines self-consistency and reward scoring, noted as SC+RM (Self-Consistency + Reward Model). Traces are grouped by their final answers, and the answer with the highest total reward is selected:

$$a_{sc+rm} = \arg\max_a \sum_{j=1}^{N} \mathbb{I}(a_{S_j} = a) \cdot \text{RM}(q, S_j),$$

where $\text{RM}(q, S_j)$ is the reward score of the j-th reasoning trace assigned by PRM for question q. A strong reward model improves selection by assigning higher scores to correct reasoning traces.

4.2 Policy Model Fine-tuning

We fine-tune the policy model using a rejection sampling guided by Med-PRM. For each question in the training set, multiple reasoning traces are generated, and Med-PRM assigns stepwise reward

scores. Traces are ranked by their minimum step score, and only top-ranked traces are retained for supervised fine-tuning.

Following Qwen et al. (2025), we exclude questions consistently answered correctly to concentrate training on more challenging examples. After fine-tuning, Med-PRM is again used at inference time to rescore multiple generations and select the best one. This bootstraps the policy model to produce reasoning paths aligned with Med-PRM, improving performance on complex medical QA tasks.

5 Experimental Setup

5.1 Training of Med-PRM

Model We perform full fine-tuning of the Llama-3.1-8B-Instruct model on a single NVIDIA A100 (80GB VRAM) with a maximum sequence length of 4096 tokens. We use the AdamW optimizer with a learning rate of 2×10^{-6}, cosine decay, and 5% warmup ratio. Training is performed in bfloat16 precision with gradient checkpointing and Flash Attention V2 with gradient accumulation to global batch size of 64. We designate the EOS token as the padding token and introduce a special marker to segment reasoning steps for process-level supervision. More details on the hyperparameters used are described in Appendix C.

Data Filtering and Labeling Training uses MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MMLU (Hendrycks et al., 2021). We use the full MedQA training set (10,178 questions) and sample 500 instances from each of the remaining datasets. For each question, we sample 16 candidate reasoning traces and filter out traces with fewer than three or more than nine reasoning steps to avoid overly shallow or degenerate reasoning. To maintain label balance, the number of correct reasoning traces per question was limited to no more than the number of incorrect traces or two, whichever is greater. To ensure the total number of tokens to not exceed 4096, 1024 tokens were reserved for the question and reasoning, and the remaining 3072 tokens were used to sequentially include relevant documents. Retrieved documents (truncated to 3072 tokens) are prepended to the question and a reasoning trace with each step separated using a special token to form the input. stepwise binary supervision is applied at each marker using labels from RAG-AS-A-JUDGE.

5.2 Evaluation of Med-PRM

Benchmarks We evaluate Med-PRM on MedQA (4 and 5 options), MedMCQA (validation set), six medical MMLU subsets (Singhal et al., 2023a), DDXPlus (Tchango et al., 2022), and two open-ended AgentClinic variants (Schmidgall et al., 2024) based on NEJM and MedQA. AgentClinic adopts an open-ended format and is evaluated using Gemini-2.0-flash. Detailed descriptions of benchmarks are in Appendix F.

Baselines We benchmark Med-PRM against proprietary and open-source models, including general-purpose, reasoning, medical, and process reward models. Table 1 summarizes performance across both multiple-choice and open-ended question answering benchmarks widely used in the medical domain.

We further compare our method against internal baselines in the ablation study: (1) PRM_{soft} and (2) PRM_{hard}, both using MCTS-style auto-labeling (Wang et al., 2024b), and (3) Med-PRM without retrieval. Our final method, (4) Med-PRM with retrieval, is also included for comparison. Each baseline is evaluated under two strategies: Best-of-N and SC+RM.

Scaling Test-Time Computation We gradually increase the number of reasoning traces generated by the policy model up to $N = 64$ per question. The final answer is selected using two strategies: Best-of-N, which chooses the trace with the highest r_{S_j} score, and SC+RM, which selects the answer group with the highest sum total reward score.

6 Results

6.1 Main Results

Table 1 summarizes the performance of Med-PRM compared to several baselines across seven medical benchmarks. We evaluate Med-PRM using two test-time strategies: Best-of-N and SC+RM, achieving average accuracy of 72.59% and 73.50%, respectively. These results outperform all existing open-source language, reasoning, medical, and medical process reward models with fewer than 10 billion parameters. With SC+RM, Med-PRM achieves state-of-the-art results on 4 out of 7 benchmarks and on 2 out of 7 benchmarks under the Best-of-N setting.

We observe larger performance improvements on benchmarks requiring complex clinical reasoning compared to knowledge-centric tasks such as

Category	Model	Size	Multiple-Choice QA					Open-Ended QA		Average
			MedQA-4	MedQA-5	MedMCQA	MMLU-Med	DDXPlus	Agent Clinic NEJM *	Agent Clinic MedQA *	
Proprietary Language Models	Gemini Flash 2.0	–	87.51	85.23	72.60	92.01	75.00	70.83	87.74	81.56
	GPT-4o-mini	–	79.03	74.31	68.20	87.79	76.00	58.33	79.44	74.73
	GPT-3.5 turbo	–	69.91	65.44	57.00	76.77	73.80	51.72	77.93	67.51
Proprietary Reasoning Models	o4-mini	–	93.95	91.12	79.60	93.99	79.80	76.67	94.86	87.14
	o3-mini	–	92.69	90.97	75.50	93.01	79.80	78.33	96.26	86.65
Open-source Language Models	Llama3.1	8B	70.93	65.20	61.60	78.97	68.80	35.83	71.96	64.76
	Gemma 2	9B	64.73	60.25	53.00	77.87	64.40	41.67	66.36	61.18
	Ministral	8B	56.17	50.43	49.20	67.22	51.80	34.17	62.62	53.09
Open-source Reasoning Models	DeepSeek-R1	671B	90.34	89.87	78.80	94.40	79.60	79.83	91.12	86.28
	QwQ	32B	85.31	81.62	70.20	88.89	74.00	63.33	85.51	78.41
	Sky-T1-Preview	32B	77.77	73.53	66.20	88.34	74.00	53.33	81.31	73.50
	R1-Distill-Llama	8B	34.96	30.16	43.60	64.19	36.80	30.83	57.01	42.51
	R1-Distill-Qwen	7B	24.82	19.56	36.40	47.47	36.80	8.33	35.51	29.84
	Sky-T1	7B	34.09	30.64	36.20	53.17	47.40	6.67	29.91	34.01
	Marco-o1	7B	39.36	34.80	49.20	69.15	38.40	30.83	63.55	46.47
Open-source Medical Models	TX-Gemma	9B	41.56	35.11	36.00	52.34	57.80	20.00	50.00	41.83
	Meditron3	8B	59.94	52.95	48.20	67.86	67.80	42.50	67.29	58.08
	Meerkat	8B	71.25	69.13	56.40	76.40	70.00	43.33	76.40	66.13
	UltraMedical	8B	72.66	68.34	62.60	79.61	72.60	45.83	70.56	67.46
	HuatuoGPT-o1	8B	72.19	63.24	63.60	75.30	64.00	40.00	71.50	64.26
Open-source Medical Process Reward Models	MedS^3									
	Best-of-N	8B	71.56	68.42	<u>64.20</u>	80.18	75.40	52.50	74.30	69.51
	SC+RM	8B	75.64	71.41	<u>64.20</u>	81.79	74.60	**55.00**	74.77	71.06
	Med-PRM									
	Best-of-N	8B	<u>76.76</u>	<u>72.43</u>	<u>64.20</u>	<u>82.37</u>	**77.80**	<u>54.17</u>	**80.37**	<u>72.59</u>
	SC+RM	8B	**79.18**	**75.49**	**67.40**	**83.29**	<u>77.20</u>	52.50	<u>79.44</u>	**73.50**

Table 1: Accuracy of proprietary and open-source models across multiple-choice and open-ended medical QA benchmarks. We use instruction-tuned models for Llama3.1, Gemma2, and Ministral. Best scores are shown in **bold**, and second-best scores are <u>underlined</u> among small-scale models ($<$ 10B parameters). We report results on AgentClinic*, simplified variants of the original benchmarks (see Appendix F for details).

MedMCQA. Notably, on the AgentClinic benchmark, which closely mirrors real-world diagnostic workflows, Med-PRM achieves accuracy gains of 12.50% and 10.75% under the SC+RM and Best-of-N settings, respectively.

Compared to MedS^3, the previous state-of-the-art process reward model at the 8B scale, Med-PRM achieves an average improvement of 2.44% across all benchmarks using the SC+RM strategy. Even under the Best-of-N strategy, Med-PRM outperforms MedS^3 by 3.08%.

These results demonstrate the strong capability of Med-PRM in identifying clinically sound reasoning paths. We further explore its effectiveness when paired with stronger, fine-tuned language models such as Meerkat-8B and UltraMedical-8B in Section 6.2

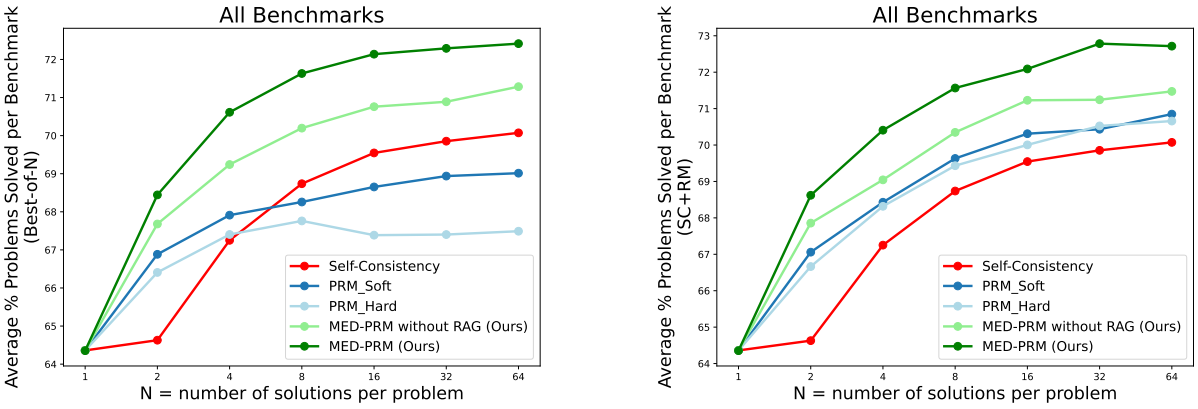
6.2 Reward Model as Verifier

To assess the model-agnostic utility of Med-PRM, we apply it as a plug-and-play verifier during inference across various policy models on MedQA. As shown in Table 2, Med-PRM consistently improves performance, regardless of the underlying base or fine-tuned model.

Model	MedQA (4 options)
Llama-3.1-8B-Instruct	68.79
+ SC	74.86 (+6.07)
+ SC + RM (Med-PRM RM)	**78.24** (+9.45)
+ Best-of-N (Med-PRM RM)	76.98 (+8.19)
Llama-3.1-8B-Instruct* (Med-PRM π)	67.22
+ SC	75.02(+7.80)
+ SC + RM (Med-PRM RM)	**79.18** (+11.96)
+ Best-of-N (Med-PRM RM)	76.76 (+9.54)
UltraMedical-8B†	67.51
+ SC	75.63 (+8.12)
+ SC + RM (Med-PRM RM)	**79.87** (+12.36)
+ Best-of-N (Med-PRM RM)	76.42 (+8.91)
Meerkat-8B†	66.65
+ SC	76.04 (+9.39)
+ SC + RM (Med-PRM RM)	**80.35** (+13.70)
+ Best-of-N (Med-PRM RM)	79.95 (+13.30)

Table 2: Performance improvements from using the Med-PRM reward model as a verifier on MedQA (4 options). For each policy model, the first row shows the average score over 64 sampled solutions. Subsequent rows apply Self-Consistency (SC), SC with reward model verification (SC+RM), and Best-of-N using the same 64 solutions.

To further demonstrate the effectiveness of Med-PRM, we train a policy model using supervised fine-tuning (SFT) on a rejection-sampled dataset constructed with our reward model, following the Entropy-Regularized PRM (Zhang et al., 2024a).



(a) Best-of-N

(b) SC+RM

Figure 2: Comparison of scaling test-time computation performance between Med-PRM and conventionally trained PRMs across overall medical benchmarks.

This policy model, denoted as Med-PRM π, leverages high-quality reasoning traces selected by Med-PRM and achieves 79.18% accuracy on MedQA with 10.39% improvement over the base Llama-3.1-8B-Instruct model alone.

Moreover, when Med-PRM is used as a verifier on top of strong, fine-tuned models such as Meerkat-8B and UltraMedical-8B, we observe additional gains. Notably, pairing Med-PRM with Meerkat-8B yields 80.35% accuracy on MedQA, marking the first time that an 8B-scale model has surpassed the 80% threshold on this benchmark. These results highlight the generalizability of our reward model as a plug-and-play component for enabling more accurate medical reasoning across diverse models. A comprehensive table of full benchmark results, including CoT, SC, and PRM baselines (hard label, soft label, MedS^3, and Med-PRM) is provided in Appendix J.

7 Analysis

7.1 Ablation Study

We conduct a comprehensive ablation study to evaluate the contribution of each component in Med-PRM. The results are visualized in Figure 2. We assess performance improvements under test-time scaling using 64 sampled solutions from the Llama-3.1-8B-Instruct model, evaluated with both Best-of-N and SC+RM strategies across the benchmark.

We compare Med-PRM to Self-Consistency and two PRMs trained with conventional auto-labeling methods: PRM_{soft} and PRM_{hard}. The results including MedS^3 are provided in Appendix G.

To analyze the impact of our design, we incrementally add key components. First, we replace conventional labeling with using an LLM to directly evaluate the reasoning steps. This is used to train a variant named "Med-PRM without RAG". Second, we incorporate retrieval for the full Med-PRM framework.

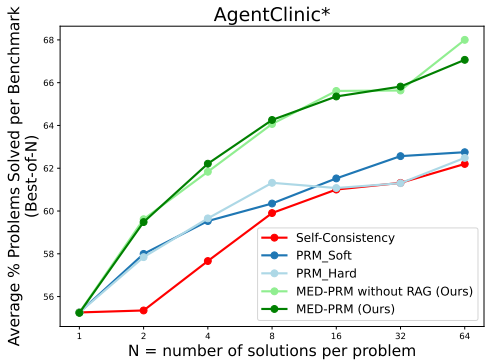


Figure 3: Comparison of scaling test-time computation performance between Med-PRM and other PRMs on AgentClinic* under the Best-of-N setting.

Results show that each component yields consistent gains across both test-time strategies. Notably, Med-PRM without retrieval already outperforms conventional PRMs, and adding retrieval further boosts performance. This demonstrates the critical role of grounding in external knowledge.

Under the SC+RM setting, conventional PRMs achieve modest improvements over Self-Consistency. However, in the more stringent Best-of-N setting, where only the top solution is selected, conventional PRMs underperform relative to Self-Consistency. In contrast, Med-PRM consistently outperforms Self-Consistency in both settings, underscoring that LLM-based step-level su-

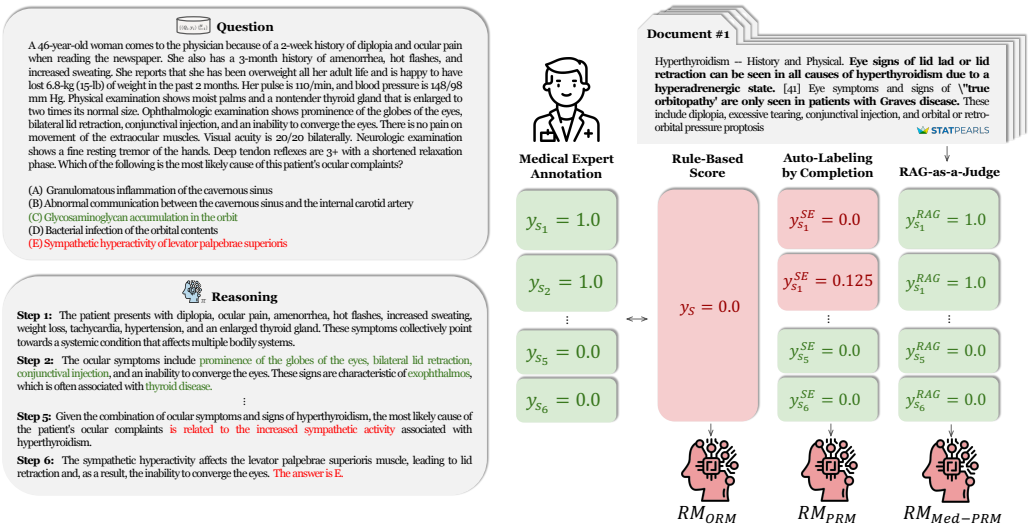


Figure 4: Case study comparison of labeling strategies for reward model training. This example illustrates how MED-PRM labeling yields more clinically accurate judgments than both rule-based and auto-completed annotations.

pervision is more effective than sampling-based auto-labeling. These findings also suggest that Best-of-N serves as a more discriminative setting for comparing reward model performance.

7.2 Open-Ended Clinical Tasks

We evaluate Med-PRM on AgentClinic* with Best-of-N strategy, a diagnostic benchmark designed in an open-ended QA format to closely resemble real-world clinical settings (Figure 3). Although this dataset was not included in the training phase, Med-PRM achieves an 11.81% improvement in accuracy through scaling test-time computation. It outperforms other baseline methods by a significant margin, achieving 4.87% higher accuracy than Self-Consistency and 4.32% higher than PRMs trained with conventional approaches.

7.3 Expert Alignment

To assess the alignment between Med-PRM and medical experts, we calculate the Pearson correlation between model-generated labels and expert annotations on step-level reasoning quality. We select three questions each from an easy subset (where Llama-3.1-8B-Instruct achieves over 10% accuracy) and a hard subset (accuracy below 10%) from the PRM training set. For each question, five model-generated reasoning traces were annotated by human experts, resulting in 180 step-level annotations in total.

As shown in Table 3, Med-PRM shows high correlation with human judgments across both easy and hard subsets (0.74 and 0.71, respectively). In contrast, the performance of soft and hard labeling strategies drops significantly on hard examples, dropping from 0.64 to 0.34 and 0.70 to 0.31, respectively. This suggests that Med-PRM produces more robust and consistent labels even in more challenging reasoning scenarios, making it better suited for constructing high-quality training sets.

Subset	Med-PRM	Soft label	Hard label
Easy	**0.74**	0.64	0.70
Hard	**0.71**	0.34	0.31

Table 3: Pearson correlation between model-generated labels and human annotations on reasoning steps for easy and hard subsets.

7.4 Case Study

Training Data Labeling Figure 4 presents an example from the MedQA training dataset concerning a patient suspected of having Graves' ophthalmopathy. Although diplopia and ocular pain are commonly associated with Graves' disease, they result from autoimmune orbitopathy rather than sympathetic overactivity.

Step 1 and Step 2 of the policy model's reasoning appropriately integrate the patient's symptoms to suspect thyroid-related exophthalmos, demonstrating sound medical reasoning. However, in Step 5, the model incorrectly attributes ocular symptoms to sympathetic overactivity, ultimately leading to the selection of an incorrect final answer.

The retrieved document clarifies that diplopia

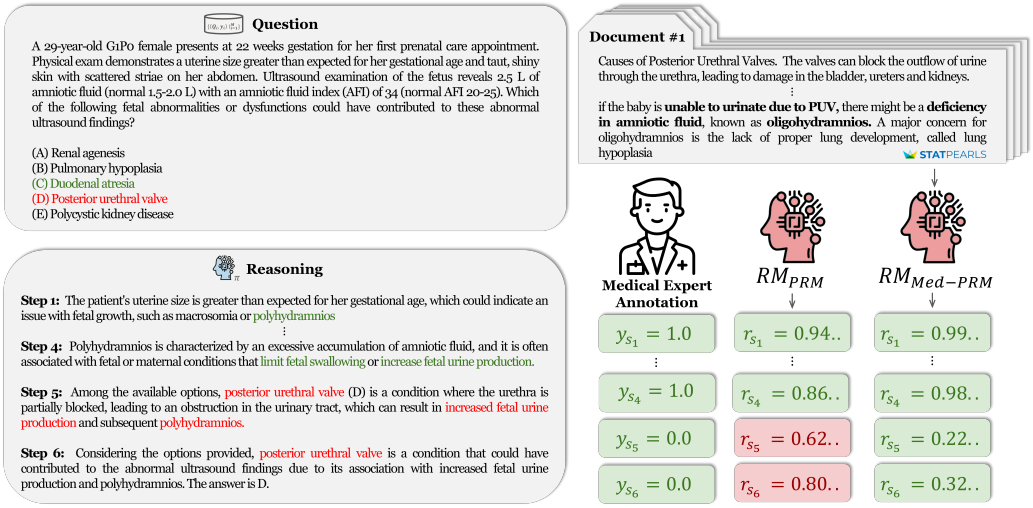


Figure 5: Case study comparison of step-wise reward assignment between conventional PRM and Med-PRM. Med-PRM assigns more appropriate and clinically grounded rewards than conventional PRM by leveraging evidence from retrieved documents, enabling more accurate evaluation of intermediate reasoning steps.

and conjunctival injection are characteristic only of autoimmune orbitopathy in Graves' disease. This distinction is key to differentiating the incorrect choice (E) from the correct one (C). Leveraging this evidence, Med-PRM assigns step-level labels of 1 to Step 1 and Step 2, and 0 to Step 5 and Step 6, aligning exactly with expert annotations.

In contrast, the ORM data labeling pipeline assigns an overall incorrect label to the entire reasoning trace, as it only considers the final answer, missing the valid reasoning steps in Steps 1 and 2. Similarly, the conventional PRM data labeling pipeline assigns low scores to Steps 1 and 2, incorrectly marking them as poor reasoning, since the final answer derived from these steps is incorrect. As a result, both ORM and PRM fail to correctly evaluate the quality of intermediate reasoning steps, mislabeling valid reasoning as low quality.

Test-Time Labeling Figure 5 presents an example from the MedQA dataset concerning a patient suspected of polyhydramnios. Duodenal atresia can cause polyhydramnios, while posterior urethral valve (PUV) typically leads to oligohydramnios.

Steps 1 to 4 of the reasoning trace appropriately integrate the patient's clinical findings to analyze the polyhydramnios context. However, in Step 5, the model incorrectly reasons that posterior uretheral valve can cause polyhydramnios, ultimately leading to an incorrect answer in Step 6.

The retrieved document explicitly states that PUV leads to oligohydramnios due to urinary outflow obstruction. This serves as a clear rationale for why the incorrect answer choice (D) is invalid.

Conventional PRM models assign relatively high rewards to Step 5 and Step 6 in the absence of helpful external documents. In contrast, Med-PRM uses the retrieved document to detect the error more precisely, providing high rewards up to Step 4 and then sharply reducing rewards from Step 5 onward.

These examples show that Med-PRM can identify and localize errors in medical reasoning steps more effectively than auto-labeling trained PRM, highlighting the value of document-grounded, step-wise evaluation in capturing nuanced reasoning quality by referring to relevant clinical documents.

8 Conclusion

In this study, we present Med-PRM, a process reward modeling framework that advances medical reasoning in language models with RAG-AS-A-JUDGE approach. Med-PRM addresses key limitations of existing PRMs by verifying intermediate reasoning steps against retrieved medical documents rather than relying on outcome-based auto-labeling. Our experiments show that Med-PRM consistently outperforms prior methods across both multiple-choice and open-ended clinical benchmarks, with notable gains in complex diagnostic tasks. Furthermore, the retrieval-based step verification produces scalable, expert-aligned labels that support effective policy training and verifier in test-time scaling. Together, these contributions establish Med-PRM as a robust and generalizable approach for improving the accuracy, transparency, and trustworthiness of medical AI systems.

Limitations

Our work has a few limitations that warrant discussion. First, our evaluation is currently confined to the medical domain, though the methodology could potentially generalize to other domains requiring stepwise reasoning verification and retrieval-augmented generation. Second, due to computational constraints, we limited our experiments to small language models like Llama 3.1 8B and Meerkat 8B, though there remains significant potential to explore scalability across different model architectures and sizes Third, while our reward model demonstrates strong performance, we did not extensively explore diverse reinforcement learning methods that could further enhance our method's capabilities. Future work should investigate these aspects through broader domain coverage, model scaling experiments, and more sophisticated reinforcement learning training strategies.

References

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. *arXiv preprint*. ArXiv:2412.18925 [cs].

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint*. ArXiv:2311.16079 [cs].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models. *arXiv preprint*. ArXiv:2404.05221 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

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 [cs] version: 2.

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 [cs].

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics (Oxford, England)*, 39(11):btad651. Publisher: Oxford University Press.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1. Publisher: Nature Publishing Group.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):1–10. Publisher: Nature Publishing Group.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. *arXiv preprint*. ArXiv:2404.15155 [cs].

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. LLMs Can Easily Learn to Reason from Demonstrations Structure, not content, is what matters! *arXiv preprint*. ArXiv:2502.07374 [cs].

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making Language Models Better Reasoners with Step-Aware Verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint*. ArXiv:2305.20050 [cs].

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023a. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023b. Med-Flamingo: a Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, pages 353–367. PMLR. ISSN: 2640-3498.

OpenAI. 2024. Introducing OpenAI o1.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. *arXiv preprint*. ArXiv:2203.14371 [cs].

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint*. ArXiv:2405.07960 [cs].

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning. *arXiv preprint*. ArXiv:2410.08146 [cs].

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023b.

Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv preprint*. ArXiv:2305.09617 [cs].

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. 2025. MedAgentsBench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning. *arXiv preprint*. ArXiv:2503.07459 [cs].

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *arXiv preprint*. ArXiv:2311.10537 [cs].

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. *arXiv preprint*. ArXiv:2205.09148 [cs].

Qwen Team. 2024. QwQ: Reflect Deeply on the Boundaries of the Unknown. Section: blog.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint*. ArXiv:2211.14275 [cs].

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025. TxGemma: Efficient and Agentic LLMs for Therapeutics. *arXiv preprint*. ArXiv:2504.06196 [cs].

Guanchu Wang, Junhao Ran, Ruixiang Tang, Chia-Yuan Chang, Chia-Yuan Chang, Yu-Neng Chuang, Zirui Liu, Vladimir Braverman, Zhandong Liu, and Xia Hu. 2024a. Assessing and Enhancing Large Language Models in Rare Disease Question-answering. *arXiv preprint*. ArXiv:2408.08422 [cs].

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, and Charles Xie. 2021. Milvus: A Purpose-Built Vector Data Management System. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, pages 2614–2627, New York, NY, USA. Association for Computing Machinery.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint*. ArXiv:2203.11171 [cs].

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. *arXiv preprint*. ArXiv:2201.11903 [cs].

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv preprint*. ArXiv:2402.13178 [cs].

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32. ArXiv:2307.14385 [cs].

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068. Publisher: Massachusetts Medical Society.

Hanning Zhang, Pengcheng Wang, Shizhe Diao, Yong Lin, Rui Pan, Hanze Dong, Dylan Zhang, Pavlo Molchanov, and Tong Zhang. 2024a. Entropy-Regularized Process Reward Model. *arXiv preprint*. ArXiv:2412.11006 [cs].

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024b. UltraMedical: Building Specialized Generalists in Biomedicine. *arXiv preprint*. ArXiv:2406.03949 [cs].

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions. *arXiv preprint*. ArXiv:2411.14405 [cs].

Qinyue Zheng, Salman Abdullah, Sam Rawal, Cyril Zakka, Sophie Ostmeier, Maximilian Purk, Eduardo Reis, Eric J. Topol, Jure Leskovec, and Michael Moor. 2025. MIRIAD: Augmenting LLMs with millions of medical query-response pairs. *arXiv preprint*. ArXiv:2506.06091 [cs].

Jiachen Zhu, Congmin Zheng, Jianghao Lin, Kounianhua Du, Ying Wen, Yong Yu, Jun Wang, and Weinan Zhang. 2025. Retrieval-Augmented Process Reward Model for Generalizable Mathematical Reasoning. *arXiv preprint*. ArXiv:2502.14361 [cs].

A Notation

Symbol	Description
q	medical question
D	retrieved documents
K	number of reasoning steps
K_j	number of reasoning steps in trace S_j
i	step index $(1, \ldots, K)$
N	number of reasoning traces
j	trace index $(1, \ldots, N)$
C	set of reasoning traces
S	reasoning trace
S_j	reasoning trace j
s_i	step i of reasoning trace S
$s_{i,j}$	step i of reasoning trace S_j
y_{S_j}	gold label for trace j
y_{s_i}	gold label for step s_i
$y_{s_i}^{SE}$	soft label for step s_i
$y_{s_i}^{HE}$	hard label for step s_i
$y_{s_i}^{RAG}$	label made by RAG-AS-A-JUDGE for step s_i
$y_{s_{i,j}}$	gold label for step $s_{i,j}$ of trace j
r_{S_j}	reward (sigmoid) score for trace j
r_{s_i}	reward (sigmoid) score for step s_i
$r_{s_{i,j}}$	reward (sigmoid) score for step $s_{i,j}$ of trace j
a^*	gold answer of q
a_{S_j}	decoded answer of trace j
a_{RM}	answer selected by reward model
$RM(D, q, S_j)$	reward score of trace j
HE	Hard Label
SE	Soft Label
SC	Self Consistency

Table 4: Summary of notations

Table 4 summarizes the notations used throughout the manuscript for the convenience of readers based on Wang et al. (2024b)

B Retrieval

We use MedCPT (Jin et al., 2023) bi-encoder and cross-encoder for dense retrieval and reranking from four biomedical corpora:

- Clinical Guidelines (Chen et al., 2023)
- StatPearls (Xiong et al., 2024)
- Medical Textbooks (Singhal et al., 2023b)
- Rare Disease Corpus (Wang et al., 2024a)

Retrieval is performed on AWS EC2 c5.9xlarge with Milvus (Wang et al., 2021) using Max Inner Product Search (MIPS); reranking uses an NVIDIA RTX 3090. For each query, 100 documents per corpus (400 total) are retrieved, with top 32 selected after reranking, following (Sohn et al., 2025).

C Hyperparameters

Hyperparameter	Reward	Policy
Learning Rate	2e-6	1e-5
Learning Rate Scheduler Type	cosine	cosine
Warmup Ratio	0.05	0.05
Batch Size	64	64
Epochs	3	1
Max Token Length	4096	4096
Precision	bfloat16	bfloat16
Optimizer	AdamW	AdamW

Table 5: Hyperparameters used for training Med-PRM

Table 5 shows the hyperparameters used for training our models of Med-PRM framework

D API Usage and Cost

We utilized the Gemini-2.0-flash model via the Google Generative AI API. The model was employed as a RAG-AS-A-JUDGE, specifically for generating training labels for the PRM and for scoring responses in the AgentClinic benchmark. All API calls were made using the official endpoint, with the temperature set to 0 and standard rate limits applied. The total API cost incurred for curating the training set of PRM was approximately $20.

E Related Works

Medical Reasoning Large language models (LLMs) have shown growing competence in medical reasoning, which presents unique challenges beyond general language tasks, including specialized terminology, multimodal patient data, and high demands for factual accuracy. Leveraging biomedical corpora such as PubMed, MIMIC-III/IV (Johnson et al., 2023), and UMLS, recent models have progressed from surface-level recall to more complex diagnostic and therapeutic inference. Pioneered by Med-PaLM (Singhal et al., 2023a), demonstrates strong performance on expert-level medical questions.

Advances in Chain-of-Thought (CoT) prompting and training (Wei et al., 2023; Xu et al., 2024; Kim et al., 2025), Retrieval-Augmented Generation (Zakka et al., 2024; Sohn et al., 2025; Zheng et al., 2025), agentic systems (Kim et al., 2024; Tang et al., 2024; Schmidgall et al., 2024; Tang et al.,

2025) further extend this paradigm by orchestrating multiple specialized agents to collaboratively solve complex clinical tasks. PRM (Lightman et al., 2023; Jiang et al., 2025), and Reinforcement Learning (Wang et al., 2024b) have enabled LLMs to generate structured reasoning traces. Similarly, HuatuoGPT-o1 (Chen et al., 2024) incorporates verifier feedback to iteratively refine multi-step reasoning traces.

These trends highlight the growing importance of process reward mechanisms. Like mathematical problem solving, clinical reasoning often requires multi-step inference, where a single incorrect step can invalidate the final outcome. Thus, stepwise supervision is not only applicable to medicine but also critical for ensuring transparency and reliability in clinical decision-making.

PRM in Medical Domain A recent study has been made to apply PRM in the medical domain. MedS^3 (Jiang et al., 2025) constructs a PRM training dataset with similar approaches to existing PRM frameworks, using MCTS-based auto-labeling. Med-PRM also provides process rewards for medical reasoning, but mainly differs in the way of constructing the training set, leveraging retrieval-augmented generation and LLM-as-a-Judge. Section 6 and Section 7 show that our approach is more effective than MCTS-based methods.

LLM-as-a-Judge for Reasoning Evaluation Recent research actively explores the use of LLM-as-a-Judge as a PRM or as a Process Explanation Model (PEM). In Hao et al. (2024), LLMs are employed to evaluate the quality of Chain-of-Thought (CoT) reasoning. Med-PRM uses RAG-AS-A-JUDGE as a labeling strategy for reasoning steps, which differs in terms of how it is utilized, and this also contrasts with the automatic labeling approaches used in prior PRM research.

Retrieval-Augmented PRM in Mathematics In the mathematical domain, a recent work has explored combining retrieval with PRM (Zhu et al., 2025). RAG-PRM retrieves semantically similar sets of questions and answers to enable PRM to generalize against out-of-distribution questions. While both Med-PRM approaches incorporate retrieval, Med-PRM differs in that it retrieves medical evidence and knowledge rather than similar QA pairs for few-shot-style prompting. Moreover, our retrieval method supports scalable integration, ranging from curated sources such as medical textbooks and clinical guidelines to potentially broader corpora like PubMed in other deployments.

F Benchmarks

Below are detailed descriptions of the medical benchmark datasets used in our study.

MedQA MedQA is a comprehensive medical question answering dataset derived from professional medical board examinations. The dataset spans three languages: English, simplified Chinese, and traditional Chinese. Our work focuses exclusively on the English subset, which contains 12,730 questions from the United States Medical Licensing Examination (USMLE). Our method was evaluated on questions with both four and five options. MedQA evaluates diverse aspects of medical knowledge, encompassing diagnostic procedures, treatment protocols, and fundamental medical concepts. The questions are designed to test both factual medical knowledge and clinical reasoning capabilities.

MedMCQA MedMCQA (Pal et al., 2022) is an extensive multiple-choice dataset comprising over 194,000 high-quality questions from Indian medical entrance examinations (AIIMS and NEET PG). Our evaluation incorporates 500 questions from this dataset, with an average token length of 12.77 tokens per question. Each question presents four answer choices. The dataset demonstrates remarkable topical diversity, covering 2,400+ healthcare topics across 21 medical subjects. MedMCQA's comprehensive coverage and focus on entrance exam questions make it particularly valuable for assessing both theoretical knowledge and practical clinical reasoning abilities in medical problem-solving scenarios.

MMLU (Medical Subset) The Massive Multitask Language Understanding benchmark (Hendrycks et al., 2021) contains specialized medical knowledge subsets that we incorporate into our evaluation. Our benchmark utilizes 1,089 medical-related questions from MMLU. Each question presents four multiple-choice options. The medical subsets encompass diverse domains including anatomy, clinical knowledge, college medicine, medical genetics, and professional medicine. MMLU's comprehensive coverage spans both fundamental and advanced medical concepts, testing knowledge across a wide spectrum of difficulty levels from basic to professional

expertise. The benchmark's standardized assessment format enables meaningful comparisons between medical reasoning capabilities and other knowledge domains.

DDXPlus DDXPlus (Tchango et al., 2022) is a large-scale synthetic dataset containing approximately 1.3 million patient cases, designed to advance research in Automatic Symptom Detection (ASD) and Automatic Diagnosis (AD) systems. Unlike traditional medical datasets that only include binary symptoms and antecedents, DDXPlus incorporates categorical and multi-choice symptoms, along with hierarchical symptom organization. Each case includes comprehensive information such as differential diagnoses, ground truth pathologies, symptoms, and relevant antecedents. This dataset enables the development of more sophisticated medical reasoning systems that can interact with patients in a logical manner and provide differential diagnoses, which is crucial for helping doctors understand the reasoning process of AI systems.

AgentClinic* AgentClinic (Schmidgall et al., 2024) is a multimodal agent benchmark designed to evaluate large language models (LLMs) in simulated clinical environments. Unlike traditional static question answering benchmarks, AgentClinic captures the complex, sequential nature of clinical decision making by integrating diverse clinical findings derived from patient interactions, multimodal data collection, and tool usage. The benchmark spans nine medical specialties and seven languages, providing a comprehensive evaluation framework. Notably, when MedQA problems are presented in AgentClinic's sequential decision making format, diagnostic accuracies can drop significantly compared to traditional formats. The benchmark enables novel patient-centric metrics and supports various tools including experiential learning, adaptive retrieval, and reflection cycles. AgentClinic's interactive environment allows for in-depth evaluation of clinical reasoning capabilities through real-world electronic health records and clinical reader studies.

In this work, we adopt a simplified variant of the benchmark, referred to as AgentClinic*, which reformulates the task into a single step inference problem. This adaptation is motivated by practical considerations: conducting multiple reasoning steps would incur excessive API calls and computational overhead in large-scale experiments. Moreover, techniques like self-consistency, which are important for evaluating model reliability, are less applicable in multi-turn settings due to non-deterministic agent trajectories. AgentClinic* thus strikes a balance between realism and tractability while preserving the core challenge of evidence-grounded clinical reasoning.

Training We train our model using four widely adopted medical datasets: MedQA, MedMCQA, PubMedQA, and MMLU-Med. For MedQA, we utilize the entire training set comprising 10,178 questions. For the remaining three datasets, we randomly sampled 500 examples from each training set to construct a lightweight yet diverse training corpus, encompassing a variety of question formats and clinical topics. This setup ensured data efficiency while enabling the model to learn from a broad range of medical problem types.

Evaluation Model performance was evaluated on MedQA, MedMCQA, PubMedQA, and six medical-related subsets of MMLU (Anatomy, Clinical Knowledge, College Biology, College Medicine, Medical Genetics, and Professional Medicine). Additionally, we conducted out-of-domain evaluations that required more complex clinical reasoning. These included DDXPlus and two variants of AgentClinic based on NEJM case reports and MedQA. For DDXPlus, we randomly sampled 500 examples due to its extensive size, and reformulated the task to focus on differential diagnosis by providing supporting evidence and requiring the model to select the correct disease from a list of up to five candidates. AgentClinic, in contrast, presented an open-ended question-answering format without predefined answer choices, simulating real-world clinical scenarios through the analysis of diverse clinical findings (multi-turn dialogues were not included). To evaluate responses in open-ended settings, we adopted an LLM-as-a-judge framework (Gemini-2.0-flash), following a similar approach to HuatuoGPT-o1 (Chen et al., 2024).

G Full Ablation Study of PRMs vs SC

As shown in Figure 6, Med-PRM outperforms MedS^3 when scaling test-time computation.

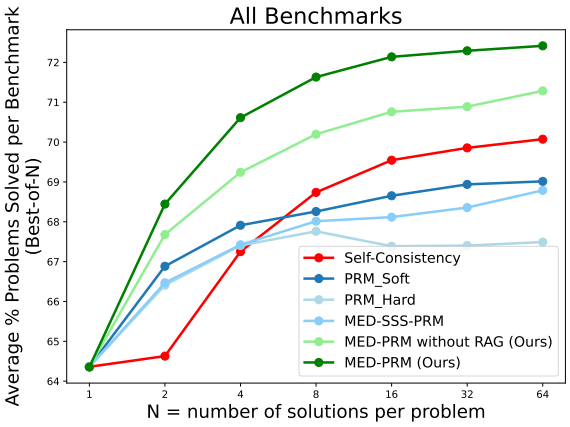


Figure 6: Comparison of scaling test-time computation performance between Med-PRM and PRMs trained with conventional approach across overall medical benchmarks(including MedS^3 in the same setting).

H Prompts

Multiple Choice Questions CoT Prompt `system_message`

Solve the following question step-by-step. Do not analyze individual options in a single step. Each step of your explanation must start with 'Step number: ' format. You must provide the answer using the phrase 'the answer is (option alphabet)' at the end of your step.

Open-Ended Questions CoT Prompt `system_message`

Solve the following question step-by-step. Each step of your explanation must start with 'Step number: ' format. The final answer must output a concise and clearly defined diagnostic term. You must provide the final answer using the phrase '## Final Diagnosis: Disease name' at the end of your final step. Please refer to the following example. ## Final Diagnosis: Multiple Sclerosis

System Message for Open-Ended Question Evaluation Using LLM-AS-A-JUDGE `system_message`

The following presents a short-answer question along with its Ground Truth and the Model's Answer. Evaluate the Model's Answer strictly based on its correctness. Your output must be either 1 or 0 only. Output 1 if the answer is correct, and 0 if it is incorrect.

Generating PRM Training Data Labels Using RAG-AS-A-JUDGE `system_message`

You are an evaluator responsible for assessing the quality of **wrong solutions** to medical questions in a stepwise manner. Each question is accompanied by relevant documents, a question, and the correct answer, and the quality of reasoning at each step must be evaluated. Give a score of 0 if the response lacks logical coherence or is not based on medical evidence, and 1 if this is not the case. Please note that if the explanation does not match the provided ground truth, it must be scored as 0. Critically assess the reasoning at each step. At the end of your evaluation, you must include a final summary of the scores in the following format:
Step 1: 0 or 1
Step 2: 0 or 1
Step 3: 0 or 1 ...

Med-PRM `system_message`

You are an evaluator assessing the logicality and validity of the reasoning in each step of the given explanation. In order to support the evaluation, the relevant documents, the question, and the explanation are provided sequentially. If the reasoning contains errors, output - after that step. If the reasoning in a step is logical and valid, output + after that step.

PRM `system_message`

You are an evaluator assessing the logicality and validity of the reasoning in each step of the given explanation. In order to support

the evaluation, the question and the explanation are provided. If the reasoning contains errors, output - after that step. If the reasoning in a step is logical and valid, output + after that step.

I Human Evaluation Details

The evaluation was conducted by one physician with four years of clinical experience and two senior medical students. Each of the two medical students independently annotated all reasoning steps, and all annotations—including those with disagreements—were subsequently reviewed and adjudicated by the physician. The evaluation followed the guidelines described below.

Human Evaluation Guidelines

The following content presents a stepwise explanation of a medical problem. Provide a critical evaluation of each step based on an integrated assessment of the following criteria.

Evaluation Criteria

- Factual Accuracy:Does the step accurately reflect established medical knowledge? Are there any inconsistencies or factual inaccuracies?

- Problem-Solving Relevance: Does the step contribute meaningfully to solving the problem? Does it avoid diverging into irrelevant or tangential reasoning?

- Logical Coherence: Is the reasoning based on appropriate medical knowledge and logically consistent within the clinical context?

Scoring Method

- Assign 1 point or 0 points to each step.

- 1 point: Awarded when the step is generally factually accurate, contributes to solving the problem, and demonstrates coherent medical reasoning.

- 0 points: Assigned when the step contains significant factual errors or involves reasoning that critically undermines the problem-solving process.

- If a step contains a critical error, any subsequent steps that rely on or are influenced by that error should also be scored as 0.

J Scaling Test-Time Computation with PRMs Across Multiple Models

Category	Model	Multiple-Choice QA					Open-Ended QA		Average
		MedQA-4	MedQA-5	MedMCQA	MMLU-Med	DDXPlus	Agent Clinic NEJM *	Agent Clinic MedQA *	
Open-source Language Models with Scaling Test-time Computation (Best-of-N)	Llama3.1								
	CoT	70.93	65.20	61.60	78.97	68.80	35.83	71.96	64.76
	SC	74.86	70.70	63.40	**81.63**	75.20	49.17	75.23	70.03
	PRM_{soft}	72.19	67.48	**64.40**	78.97	74.40	51.67	73.83	68.99
	PRM_{hard}	71.09	67.48	61.80	76.49	70.60	48.33	76.64	67.49
	$MedS^3$	70.15	65.99	62.80	77.78	77.60	49.17	**78.04**	68.79
	Med-PRM	**76.98**	**73.06**	63.40	81.18	**78.00**	**57.50**	76.64	**72.39**
	UltraMedical								
	CoT	72.66	68.34	62.60	79.61	72.60	45.83	70.56	67.46
	SC	75.63	71.80	64.20	81.91	**76.20**	49.17	76.64	70.79
	PRM_{soft}	75.65	71.80	63.00	81.36	73.60	**55.00**	**77.57**	71.14
	PRM_{hard}	76.28	71.48	63.40	80.99	72.80	48.33	74.77	69.72
	$MedS^3$	74.00	68.66	63.00	80.62	75.80	48.33	74.77	69.31
	Med-PRM	**76.42**	**74.94**	**64.40**	**83.20**	75.40	**55.00**	77.10	**72.35**
	Med-PRM π								
	CoT	67.16	64.26	57.20	75.48	67.20	40.00	68.69	62.86
	SC	76.04	71.80	62.20	82.00	74.80	50.83	78.04	70.82
	PRM_{soft}	72.58	68.81	60.20	78.97	74.60	49.17	77.10	68.78
	PRM_{hard}	72.27	69.05	60.40	78.33	71.40	45.83	73.83	67.30
	$MedS^3$	69.60	67.09	60.00	78.79	74.80	48.33	71.50	67.16
	Med-PRM	**76.76**	**73.06**	**64.40**	**82.46**	**77.80**	**54.17**	**80.37**	**72.72**
Open-source Language Models with Scaling Test-time Computation (SC+RM)	Llama3.1								
	CoT	70.93	65.20	61.60	78.97	68.80	35.83	71.96	64.76
	SC	74.86	70.70	63.40	81.63	75.20	49.17	75.23	70.03
	PRM_{soft}	75.49	72.19	65.00	82.28	75.00	50.83	76.17	70.99
	PRM_{hard}	75.33	71.09	64.20	81.73	75.20	51.67	75.23	70.64
	$MedS^3$	73.84	70.15	64.40	81.36	75.80	**54.17**	76.64	70.91
	Med-PRM	**78.24**	**73.53**	**66.40**	**83.29**	**76.80**	53.33	**77.10**	**72.67**
	UltraMedical								
	CoT	72.66	68.34	62.60	79.61	72.60	45.83	70.56	67.46
	SC	75.63	71.80	64.20	81.91	76.20	49.17	76.64	70.79
	PRM_{soft}	77.14	72.43	65.40	82.46	75.60	51.67	**78.50**	71.89
	PRM_{hard}	76.90	72.19	64.40	82.37	75.80	50.00	78.04	71.39
	$MedS^3$	75.41	72.03	63.00	81.36	76.60	**54.17**	75.23	71.11
	Med-PRM	**79.87**	**75.26**	**65.50**	**82.83**	**77.60**	52.50	77.57	**73.02**
	Med-PRM π								
	CoT	67.16	64.26	57.20	75.48	67.20	40.00	68.69	62.86
	SC	76.04	71.80	62.20	82.00	74.80	50.83	78.04	70.82
	PRM_{soft}	76.51	72.51	63.60	82.55	75.80	**54.17**	78.97	72.02
	PRM_{hard}	77.06	72.58	63.00	82.19	75.60	53.33	**79.44**	71.89
	$MedS^3$	75.26	72.03	63.40	81.82	75.60	52.50	78.04	71.24
	Med-PRM	**79.18**	**75.49**	**67.40**	**83.29**	**77.20**	52.50	**79.44**	**73.50**

Table 6: Accuracy of open-source language models with scaling test-time computation. Three models were evaluated: Llama 3.1 8B Instruct, Llama-3-8B-UltraMedical (best performing model below 10B parameters), and Med-PRM π, for solution sampling. The sampling outputs were assessed using Self-Consistency (SC) and various PRM methods under both Best-of-N and SC+RM settings. Across different scaling test-time computation strategies for each sampling, the best scores are shown in **bold**, and second-best scores are underlined. Med-PRM achieved the highest average score across test-time computation scaling methods.