MedS^3: Towards Medical Small Language Models with Self-Evolved Slow Thinking

Anonymous ACL submission

Abstract

Medical language models (MLMs) have become pivotal in advancing medical natural language processing. However, prior models that rely on pre-training or supervised fine-tuning often exhibit low data efficiency and limited practicality in real-world clinical applications. While OpenAI's o1 highlights test-time scaling in mathematics, attempts to replicate this approach in medicine typically distill responses from GPT-series models to open-source models, focusing primarily on multiple-choice tasks. This strategy, though straightforward, neglects critical concerns like data privacy and realistic deployment in clinical settings. In this work, we present a deployable, small-scale medical reasoning system, MedS^3, designed for long-chain reasoning in clinical tasks using a self-evolution paradigm. Starting with a seed dataset of around 8,000 instances spanning five domains and 16 datasets, we prompt a base policy model to perform Monte Carlo Tree Search (MCTS) to construct rule-verifiable reasoning chains. Each reasoning step is assigned an evolution rollout value, allowing verified trajectories to train the policy model and the process reward model (PRM). During inference, the policy model generates multiple responses, and the reward model selects the one with a newly proposed PRM-guided Vote-Sum (P-VS) strategy. Experiments on eleven evaluation datasets demonstrate that MedS^3 outperforms not only the prior strongest medical model by 6.59, but also 32B-level general reasoning models by 8.71 points.

1 Introduction

Large Language Models (LLMs) have demonstrated significant potential in the medical domain (Singhal et al., 2023; Nori et al., 2023; Chen et al., 2023b), with applications ranging from generating clinical notes (Biswas and Talukdar, 2024; Jung et al., 2024) to supporting patient communication (Tu et al., 2024; Liao et al., 2024b). Recently, slow-thinking reasoning models, exemplified by OpenAI o1 (OpenAI, 2024), have shown impressive improvements on reasoning-intensive mathematical problems (Lyu et al., 2025; Wang et al., 2024). However, limited efforts have been made to induce similarly strong reasoning abilities in medical-oriented language models.

Before the rise of OpenAI o1, researchers focused on improving medical language models (MLMs) primarily through extensive pre-training, which demands large computational resources and often yields only modest gains in downstream tasks. For instance, MMed-Llama3 post-pretrained Llama3 8B on 25.5B tokens but achieved only about a five-point performance gain compared to its base model (Qiu et al., 2024). Other works (Christophe et al., 2024; Ankit Pal, 2024) leverage supervised fine-tuning (SFT; Ouyang et al. (2022)) since it is computationally efficient. However, many human-annotated SFT datasets provide only concise responses (either short phrases or a single ground truth option), and hence fine-tuning on such data can degrade an MLM's broader language fluency, reducing its practical value in realistic clinical scenarios. On the other hand, synthetic corpora generated by LLMs (Luo et al., 2024; Qiu et al., 2024) often contain hallucinations (Xu et al., 2024b; Huang et al., 2023), so directly using these outputs for teacher-forcing limits a target model's optimization space. Consequently, focusing on inference-time scaling, often referred to as "slow thinking", emerges as a data-efficient approach with the potential to enhance clinical mastery while mitigating the drawbacks of pre-training- or fine-tuning-heavy strategies.

A closely related work to slow thinking is HuatuoGPT-o1 (Chen et al., 2024), which employs GPT-4o to generate reasoning-intensive problems and corresponding complex reasoning steps for distillation and reinforcement learning (RL). While it achieves certain levels of long-chain reasoning, its

Models	Without Pretraining	Without Close-sourced Teacher	Diverse Clinical Coverage	Small Size	Slow Thinking	Process Reward Usage
MMed-Llama3	✗	✗	✓	✓	✗	✗
UltraMedical	✓	✗	✓	✓	✗	✗
HuatuoGPT-o1	✓	✗	✗	✓	✓	✗
O1-journey Part 3	✓	✗	✗	✗	✓	✗
MedS^3	✓	✓	✓	✓	✓	✓

Table 1: Comparison of MedS^3 with other medical models. Our MedS^3 supports flexible inference-time scaling on resource-constrained devices, as well as process reward-guided decoding algorithms. Furthermore, MedS^3 is a self-evolved model without dependence on large proprietary models for distillation or critique.

heavy reliance on large proprietary models limits its generalizability to other clinical applications. Moreover, The adopted RL algorithm (Schulman et al., 2017) requires accurate value models, which is far from practicability in medicine yet.

Another relevant work, O1 Journey Part 3 (Huang et al., 2025), directly distills OpenAI o1's outputs into ∼70B-parameter models. This approach compromises user-friendliness and data privacy, and also struggles to detect hallucinations. Moreover, relying exclusively on multiple-choice problems for distillation constrains its applicability to a broader range of clinical tasks.

In contrast, this work introduces MedS^3, a small-scale medical language model equipped with robust long-chain reasoning capabilities (the "policy") and an accompanying fine-grained process reward model (the "PRM"). Our approach uses a self-bootstrapping pipeline to enhance the model's performance across diverse clinical tasks. Specifically, we curate 16 medical tasks from established medical training corpora, encompassing clinical diagnosis QA, medical natural language inference, knowledge-intensive QA, long-context QA, and biomedical QA. We then sample 500 seed instances from each task to form an initial dataset of approximately 8,000 instances, which kickstarts our self-evolution process. Leveraging a Monte-Carlo Tree Search (MCTS)-based evolution pipeline, we simultaneously generate synthetic datasets for both policy fine-tuning and PRM training. Rule-verified synthetic trajectories are preserved to optimize the policy model, where intermediate reasoning steps labeled with MCTS rollout values serve as the PRM dataset. By fine-tuning the base model on this enriched policy dataset and further enhancing it using our proposed PRM-guided Vote-Sum strategy, our system achieves holistically superior performance across eleven clinical reasoning benchmarks, surpassing both medical models and general reasoning models with much lower parameters.

The uniqueness of MedS^3 is further underscored by a comprehensive comparison in Table 1, which demonstrates its superiority over other medical models in robust long-chain reasoning and breadth of clinical-task coverage. Overall, we summarize our contributions as follows:

1. **First Self-Evolution Framework:** To the best of our knowledge, this work introduces the first self-evolution framework specifically designed to empower small-scale medical models with long-chain reasoning capabilities, enabling data-efficient performance gains across a wide range of clinical applications.

2. **State-of-the-Art Performance:** Our self-evolved system, MedS^3, achieves comprehensive performance improvements across eleven clinical benchmarks, outperforming all medical and large-sized reasoning models overall. This is driven by the integration of a fine-grained PRM that enhances reasoning accuracy at each step.

3. **Open-Source Resources for Research:** We openly release both the policy fine-tuning corpus and process reward model corpus, providing valuable resources for future research and fostering further advancements in medical AI.

2 MedS^3

This section presents a detailed overview of the proposed MedS^3 framework, which is structured into four key components:

1. **Self-Bootstrapping Evolution** (§2.1) which synthesizes reasoning trajectories as training data, with Monte-Carlo Tree Search (MCTS) technique using the base policy π_0.

2. **Policy Model** π (§2.2) which is derived by fine-tuning the base policy π_0 using the generated synthetic data with supervised learning.

3. **Process Reward Model (PRM)** V_θ (§2.3) which is fine-tuned with step-wise supervision using soft-labels and assigns a value in the range $[0, 1]$ to each reasoning step, representing correctness of individual steps.

4. **PRM-guided Inference** (§2.4) which utilizes the proposed PRM-guided Vote-Sum strategy to choose the final solution.

The overall framework is presented in Fig. 1.

2.1 MCTS-guided Evolution

This algorithm builds upon an n-ary tree, where each tree node T contains the following attributes: (1) reasoning step s, which is an intermediate step tracing from the root; (2) value v, which is an evaluation of the correctness of the current node; (3) the children nodes $\{c\}$, which is a collection of nodes that continue reasoning from the current node; (4) the parent node p which is the former reasoning step and (5) the number of visits n. Every root node is initialized as $T = ([s_0], 0, \emptyset, \text{null}, 0)$ where s_0 ="Let s break down this problem step by step." to guarantee a multi-step reasoning process. There are four stages in a full MCTS pipeline, including *Node Selection*, *Node Expansion*, *Node Rollout*, and *Backpropagation*.

Node Selection This process starts from an initial root to select the next node. Within each iteration, we use UCB (Winands et al., 2008) as the criterion to select a child, which is as follows:

$$UCB(T) = v_C + \gamma\sqrt{\frac{\ln n_{T_{parent}}}{n_T}} \qquad (1)$$

where T_{parent} is the parent node of T and γ is an exploration constant set as 2. For each intermediate node, we select its child node with the highest UCB value. We choose this criterion to expect models to further explore those nodes with high rollout values but low visiting counts, which inhibits models from repeatedly expanding already high-value nodes and encourages the expansion of underexplored nodes with the second highest value.

Node Expansion After picking up the candidate node T_c using the UCB criterion, we expand the reasoning steps of the current node. If the current node possesses a relatively high value ($v_c \geq thr$, where $thr = 0.9$ is a pre-defined threshold), we prompt the node to directly generate a `Finish` node to accomplish this path reasoning. This manual operation not only does not impact the reasoning correctness as a value is close to 1 only when the trace $[s_0, s_1, \cdots, s_k]$ is close to the correct final answer, but also reduces unnecessary exploration tokens. Otherwise, assume that the selected node is located at k-th depth among the tree with previous reasoning trajectories $[s_0, s_1, \cdots, s_k]$ connected by a coherence phrase t_s, we sample B single-step outputs $\{s_{k+1,i} \mid i = 1, 2, \cdots, B\}$ based on the previous trajectory using a `Reason`[1] node:

$$s_{k+1,i} \sim \pi_0([s_0 \oplus s_1 \oplus \cdots \oplus s_k] \mid x) \qquad (2)$$

where $\oplus$ is the operation to connect two steps using the coherence phrase t_s, π_0 is the base policy model, x is the original input prompt and $s_{k+1,i}$ is sampled with a high temperature (1.0) to harvest a diverse search space. To control that each reasoning step is atomic enough for the whole thought path, we set the stop tokens as `Step k+2`. Subsequently, the B nodes $\{T_i = ([s_0, s_1, \cdots, s_k, s_{k+1,i}], 0, \emptyset, T_c, 0) \mid i = 1, 2, \cdots, B\}$ are added to T_c as children nodes.

Node Rollout As the PRM is not yet available during evolution, the rollout process is conducted using simulation to obtain the estimated value for a chosen node. Specifically, for a chosen unvisited node T_c, we set a simulation budget $L = \min(L_{\min}, \frac{L_0}{k})$ where k is reasoning step counts, to encourage sufficient simulation trials when the known reasoning path is short, but expect to see a deterministic reasoning result conditioning on a long trajectory. After that, we prompt the policy model π_0 to directly output the answer L times:

$$a_c^l \sim \pi(x, [s_0, s_1, \cdots, s_k]) \qquad (3)$$

where $l \in [1, L]$ and a_c^l is the l-th simulated answer. The accuracy of the L simulations $acc = \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{a_c^l = y}$ is assigned as the value of T_c. For the `Finish` node, the rollout value is computed by comparing the derived answer and the ground truth and an $0/1$ value is assigned.

Backpropagation After obtaining the value for the selected node, we conduct value backpropagation starting from T_c till the root node, updating all tree node values among the trace. This

[1] Prompts of `Finish` and `Reason` actions are illustrated in Appendix D

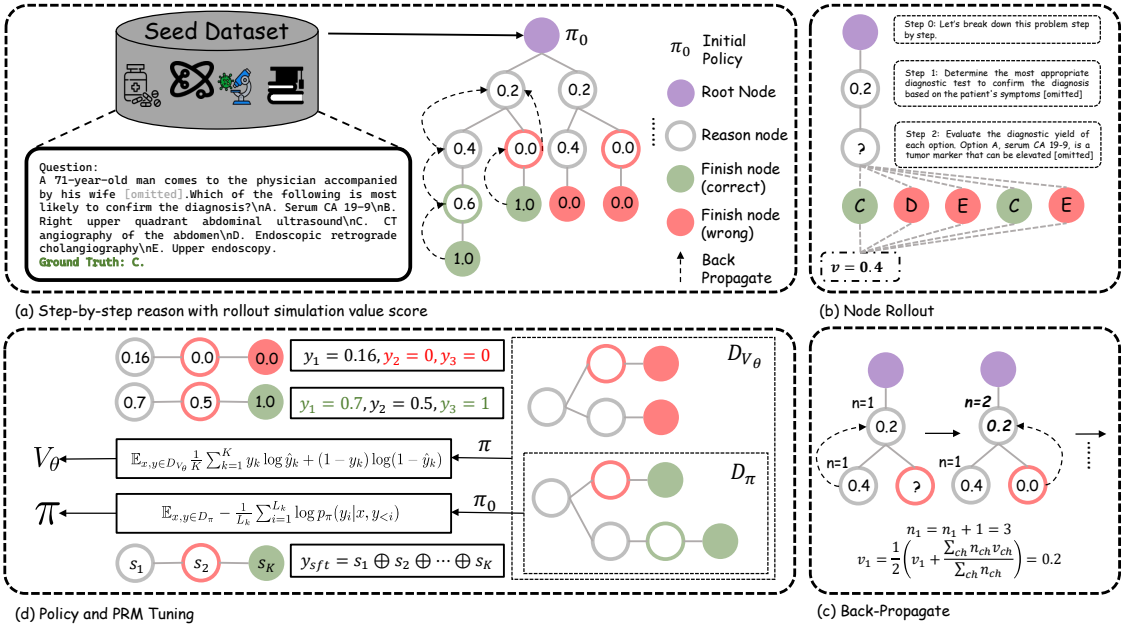

Figure 1: Overview of the construction of MedS^3 framework. MedS^3 utilizes a Monte-Carlo Tree Search pipeline to self-generate step-by-step reasoning paths for each question in the seed dataset (a). During this process, MedS^3 uses result simulation to obtain the rollout value for each node (b); After obtaining the child's rollout value, MedS^3 executes back-propagate to enable precise value prediction from deeper layer to transfer back to shallow nodes (c). After gathering all correct and wrong finish nodes, we use supervised fine-tuning to optimize the policy model π with correct reasoning trajectories and step-wise discriminative loss to obtain a process reward model V_θ (d).

process aims to use more accurate value estimation in the deeper layer to update early nodes' values so that these values serving as PRM's optimization labels become synchronously precise. Specifically, for an arbitrary node T_k, we propose to update its visits n_k and v_k as follows:

$$n_k = n_k + 1 \tag{4}$$

$$v_k = \frac{1}{2}\left(v_k + \frac{\sum_{ch} v_{ch} \cdot n_{ch}}{\sum_{ch} n_{ch}}\right) \tag{5}$$

Note that this update process considers both its children values and its own value, which guarantees that the parent node will never have the same value as its children. It considers both correctness and completeness for the evaluation of a reasoning step.

Termination of Search Due to computational limits, we cannot afford excessive computing resources to fully expand the tree for a training instance. Therefore, we set two criteria to terminate the search process. Firstly, we set the minimum correct nodes to τ. Once the total correct count in the tree exceeds τ, we stop the exploration of this tree. Second, if there are no correct nodes after affording a certain number of node exploration trials, we prompt π_0 to generate `Finish` node for all leaves. The first strategy aims to reduce extra computation, while the second expects to obtain as more correct answers as possible, for the optimization of both policy and reward models.

2.2 Policy Model Fine-tuning

The policy training mainly leverages the correct nodes T_k^1 which are assigned a value 1.0 during rollout and corresponding reasoning trajectories gathered before: $D_\pi = \{(T_k^1, [s_0 \oplus s_1 \oplus \cdots \oplus s_k])\}$. These correct reasoning traces are supervised fine-tuned to deduce a self-improved policy model:

$$\mathcal{L}_\pi = \frac{1}{L_k}\sum_{i=1}^{L_k} -\log p_\pi(y_i|x, y_{<i}) \tag{6}$$

where y_i is the i-th token of the reason trajectory and L_k is the total length of the trajectory.

2.3 PRM Fine-tuning

Dataset Collection We gather all finish nodes and their corresponding traces: $\{(T_k^j, [s_0, s_1, \cdots, s_k^j], v_k^j) \mid k = 1, 2, \cdots, K; j = 1, 2 \cdots, J\}$, where K is the max depth of the evolved tree, J is the max width of the j-th

layer, and $v_k^j \in (0,1)$. The reasoning traces $[s_0, s_1, \cdots, s_k^j]$ are assembled with t_s. This naturally enables V_θ to distinguish between siblings inheriting from the same parent node: $T_k^{j_1}$ and $T_k^{j_2}$ with distinct rollout values $v_k^{j_1}$ and $v_k^{j_2}$. To prevent V_θ from being biased by the distribution of value 0 and value 1, we tallied the correct and incorrect finish nodes: $\{T_k^0 \mid v_{T_k} = 0\}$ and $\{T_k^1 \mid v_{T_k} = 1\}$, and randomly sampled elements $\min\left(|\{T_k^0\}|, |\{T_k^1\}|\right)$ from these two sets, maintaining a balanced distribution of the nodes of correct and incorrect reasoning: $\{\hat{T}_k^0\}$ and $\{\hat{T}_k^1\}$. The sampled nodes are combined and form the final V_θ tuning set: $D_{V_\theta} = \{\hat{T}_k^0\} \cup \{\hat{T}_k^1\}$.

Learning objective Previous works in the math domain choose to directly learn the rollout value (Zhang et al., 2024a) or learn the pair-wise ranking preference (Guan et al., 2025). However, in our work, we choose to learn the prediction of the correctness probability of an intermediate step using a 2-class cross-entropy loss. The PRM V_θ is tuned based on the tuned policy model, with the language model head replaced by a token classification layer with a cross-entropy loss at the end of each step. Although Zhang et al. (2025) suggests that the PRM label should be set to 1 (a hard label) once the rollout score is above zero, empirically we borrow the insight from label-smoothing (Szegedy et al., 2016), where a soft label groups positive internal steps tightly in the representation space. We show the comparison between the soft and hard label in Appendix B.1. Specifically, we set the label y_k as the original node rollout value, and optimize V_θ using the following loss function:

$$\mathcal{L}_{V_\theta} = \frac{1}{|D_{V_\theta}|} \sum_{T_k \in D_{V_\theta}} y_k \log \hat{y}_k + (1 - y_k) \log(1 - \hat{y}_k) \tag{7}$$

where $\hat{y}_k$ is the predicted probability of the given step and y_k is the label. This soft-label training, not only encourages V_θ to cluster preferred and dispreferred steps, but also prevents the learning of fuzzy labels (rollout value around 0.5).

2.4 PRM-guided Decoding

Referring to the inference scaling law (Wu et al., 2024d), enlarging the inference token budget is a deterministic way to enhance the downstream reasoning performance. To balance between the advantages of enlarging the inference budget and unwillingly reduced efficiency, we propose a new decoding strategy named PRM guided Vote-Sum that considers both response estimation and semantic coherence, which is illustrated in Fig. 2.

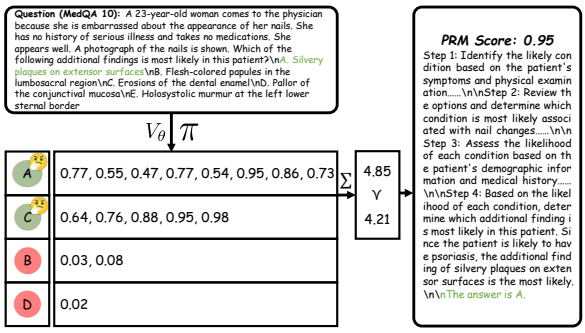

Figure 2: Overview of the PRM-guided Vote-Sum.

PRM guided Vote-Sum (P-VS) This method employs π to sample M candidate responses $\{y_m \mid m = 1, 2, \cdots, M\}$ given an input problem x and uses the PRM V_θ to select the response whose answer is estimated to have the highest values in total. Specifically, for a simple output y_m, we split it into K steps $\{s_m^k \mid k = 1, 2, \cdots, K\}$ with pre-defined coherence phrase t_s. The PRM assigns a score $q \in [0, 1]$ for each internal step $\{q_m^k = V_\theta(s_m^1 \oplus s_m^2 \oplus \cdots \oplus s_m^k) \mid k = 1, 2, \cdots, K\}$. The overall response score v_{y_m} takes the minimum value of the score chain (Lightman et al., 2023) or the last value (Zhang et al., 2025). The Vote-Sum strategy comprehensively considers the occurrence of semantically equivalent outputs but also the confidence score predicted by V_θ:

$$a_v = \{(a_1, v_{a_1}), (a_2, v_{a_2}), \cdots, (a_n, v_{a_n})\} \tag{8}$$

$$y_{target} = \underset{y_m \in \{y_j \mid a_{y_j} = \text{argmax}_{a_i} v_{a_i}, (a_i, v_{a_i}) \in a_v\}}{\text{argmax}} v_{y_m} \tag{9}$$

where $n \leq M$ is the number of different answers obtained from the M responses, $v_{a_n} = \sum_m \mathbb{1}_{a_{y_m} = a_n} \cdot v_{y_m}$ is the sum of values of samples whose answers equal to a_n.

3 Data Statistics

A slow-thinking system in medical scenarios should both excel at exam-level question answering (QA) and handling real-world clinical scenarios, like diagnosis (Tchango et al., 2022), specific disease syndrome (Lab, 2020) and drug-related problems (Huynh et al., 2016). However, previous

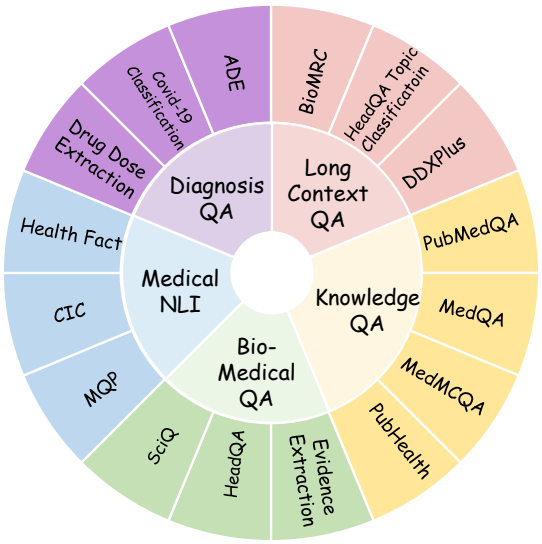

Figure 3: Overview of the used seed datasets.

works only focused on a simple scenario, with only limited data diversity, especially multiple-choice QA, to train reasoning models. To approximate realistic clinical usage and promote medical reasoning models on a diverse range of clinical tasks, we curate a training corpus, from 16 existing public medical datasets and divide them into five dimensions according to the task category. For each dataset, we randomly select 500 items and form a seed dataset with around 8,000 instances. We show the dataset statistics in Fig. 3. The details about the definition of the five dimensions and the corresponding tasks can be found in Appendix E.

4 Experiments

In this section, we comprehensively evaluate the proposed MedS[3] on various downstream domains, including both in-domain and out-of-domain datasets.

4.1 Experiment Setups

Training and Evaluation We choose Llama3.1-8B-Instruct as the initialization of both the policy model and PRM model. We select MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), PubHealth (Kotonya and Toni, 2020), BioMRC (Pappas et al., 2020), HealFact Classification (Kotonya and Toni, 2020), Drug Dose Extraction (Huynh et al., 2016), DDXPlus (Tchango et al., 2022) as the in-domain evaluation benchmarks, the medical parts of MMLU (Hendrycks et al., 2021), BioASQ (Tsatsaronis et al., 2012) and SEER Classification (Dubey et al., 2023) as the out-of-domain evaluation sets. We provide three different decoding strategies for MedS[3], including CoT (Wei et al., 2022), Self-Consistency (SC) and our proposed P-VS. The hyperparameters of synthesis, self-training and evaluation are presented in Appendix F.

Baselines We choose the following categories to serve as baselines: (1) Proprietary general LLMs, including GPT-3.5-turbo (OpenAI, 2022) and GPT-4o-mini (OpenAI, 2023); (2) Open-sourced general LLMs, including Llama 3 8B, Llama 3.1 8B (Dubey et al., 2024) and Qwen2.5 7B (Yang et al., 2024), QWQ-preview-32B (Team, 2024) and two distilled models from DeepSeek R1 (Guo et al., 2025): R1-Distill-Llama8B and R1-Distill-Qwen32B; (3) Open-sourced Medical LLMs, including MedLlama 3 8B[2], MMedS-Ins-Llama-3-8B (Wu et al., 2024b), Med42 (Christophe et al., 2024), OpenBioLLM (Ankit Pal, 2024), and UltraMedical3-8B and UltraMedical3.1-8B (Zhang et al., 2024b). We also compare our method with HuatuoGPT-o1-8B (Chen et al., 2024).

4.2 Main Results

We present the comprehensive experiment results in Table 2. The results unveil that most prior medical LLMs show superior results in traditional medical benchmarks (MedQA or PubMedQA); while such superiority cannot generalize to out-of-distribution real-world clinical benchmarks (DDXPlus or DrugDose), which results in their sub-optimal overall performance compared to Llama3-8B. In contrast, our MedS[3] is not optimized exclusively for multiple-choice medical datasets and hence achieves the best performance among all open-sourced competitions. As an 8B system, MedS[3] achieves +13.07 average performance gains with respect to the base model in the overall assessment, which not only outperforms medical-oriented models as well as general reasoning models.

Specifically, compared to HuatuoGPT-o1 and MedLlama3, MedS[3] shows superior performance on reasoning-intensive benchmarks, including MedQA and MedMCQA, as well as clinical benchmarks. This verifies that MedS[3] learns medical reasoning philosophies and the clinical deduction process. Another model MMedS-Ins, which directly post-pretrains and fine-tunes on millions of clinical corpus, harvests superior performance on

[2] https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20

Models	Knowledge QA		BioMedical QA			Long Context QA		NLI		Diagnosis QA		Avg.
	MedQA	MedMCQA	PubMedQA	BioASQ	MMLU	BioMRC	DDXPlus	PubHealth	HealthFact	DrugDose	SEER	
Proprietary language models												
GPT-4o-mini	75.81	67.58	47.80	83.01	83.79	66.85	54.00	59.14	65.24	73.91	54.54	66.52
GPT-3.5-turbo	59.31	58.12	37.40	74.11	71.11	56.22	39.05	57.84	67.85	86.96	73.61	61.96
Open-source language models												
Qwen2.5-7B	55.54	54.12	53.40	73.62	74.38	56.48	31.25	57.11	52.69	60.87	33.07	54.78
Llama3-8B	57.50	55.92	56.40	75.73	68.55	56.50	35.30	64.09	70.88	73.91	47.07	60.17
Llama3.1-8B	61.51	57.42	59.00	71.36	72.52	55.60	19.00	61.82	63.97	73.91	52.62	58.98
R1-Distill-Llama8B	50.12	48.89	46.60	70.55	68.42	53.49	36.10	55.73	62.04	69.57	31.71	53.93
QwQ-32B-preview	68.89	61.03	48.60	73.62	74.18	79.76	45.40	63.36	66.08	39.13	37.26	59.76
R1-Distill-Qwen32B	76.83	66.27	38.20	78.32	85.07	78.66	53.90	59.95	63.80	82.61	26.22	64.53
Open-source medical models												
MMedS-Ins	53.57	48.24	56.60	77.35	50.86	31.47	97.53	54.26	69.64	95.65	97.93	66.65
MedLlama3	55.85	59.36	66.40	84.63	70.08	47.97	22.50	62.39	68.10	69.57	50.69	59.78
Med42	50.20	49.70	55.40	74.76	61.43	57.26	31.35	59.14	81.57	65.22	37.14	56.65
OpenBioLLM	50.20	50.56	41.40	47.73	61.69	27.46	16.55	18.77	53.28	34.78	46.48	40.81
UltraMedical3-8B	68.89	61.82	51.60	80.58	75.08	45.18	36.70	66.13	72.73	60.87	24.55	58.56
UltraMedical3.1-8B	70.93	62.78	56.40	77.18	76.43	54.26	31.55	59.14	70.20	56.52	45.86	60.11
Open-source slow-thinking medical models												
HuatuoGPT-o1	62.53	59.31	69.20	87.70	70.53	50.98	40.20	24.61	66.08	56.52	46.85	57.68
MedS^3 (Ours)												
CoT	65.91	60.55	56.80	78.48	75.66	55.84	51.65	57.03	64.73	73.91	48.97	62.68
SC	70.93	64.21	58.20	79.13	79.63	63.66	57.00	64.42	70.37	86.96	52.19	<u>67.88</u>
P-VS	71.88	65.20	59.60	80.10	79.50	77.12	65.20	73.03	79.97	95.65	58.36	**73.24**

Table 2: Experiment results in 11 medical datasets among four types of models. We highlight the best results with **bold** and <u>underlines</u> the second-best results. MedS^3 with PRM guided Vote Sum (P-VS) achieves superior performances on real-world clinical datasets.

in-domain clinical tasks, such as SEER and DDXPlus. However, directly fine-tuning on question-answer pairs inevitably makes the model lose the ability to output long responses, which is extremely important in reasoning tasks. In contrast, our model MedS^3, possesses a comprehensive performance on both traditional tasks and clinical scenarios with strong reasoning abilities.

5 Analysis

5.1 Effectiveness of P-VS

In this section, we compare our proposed decoding method P-VS to the previously widely adopted PRM decoding method: Best-of-N (BoN) as well as ORM guided methods (ORM with Last and ORM with Vote-Sum). We compare three variants of BoN, including BoN-min (Lightman et al., 2023) which takes the minimum step value of a reasoning trajectory as the estimation of the whole sequence, BoN-prod (Lightman et al., 2023) which takes the production of step values instead, and BoN-last (Zhang et al., 2025), which treats the PRM as an ORM. We select two traditional QA tasks and three clinical tasks to conduct comparison and show results in Table 3. A very interesting finding is that in the two traditional QA datasets, BoN-last achieves extremely bad results, but shows great performance on par with Vote-Sum in clinical tasks. We hypothesize that clinical tasks and Olympia-level math problems are both difficult for Llama3.1-8B, and hence the findings that PRM behaves more like an ORM transfer successfully to clinical tasks. On the other hand, on relatively advantageous tasks, MC-rollout value encodes values of both future and current steps, which leads the PRM to perform as a human-labeled PRM to distinguish steps precisely. As a result, in traditional medical tasks, we use the minimum value among step values to represent the whole sequence's correctness, while in clinical tasks, we use the last step's value instead. Compared to ORM-based methods, they only excel at traditional QA benchmarks, lagging far behind all PRM-based methods in clinical usage, further validating the efficacy of process supervision. ORM with Vote-Sum surpasses ORM with last by a large margin, which also verifies the effectiveness of the proposed Vote-Sum strategy. We present the error rate of PRM in Appendix B.2 to show that rule-verified PRM is sufficient in detecting erroneous clinical reasoning steps.

5.2 Inference-Time Scaling

In this section, we discuss whether MedS^3 can benefit from more inference tokens. We sample $n = 2, 4, 8, 12, 16$ candidates for a given problem, and conduct P-VS to select the most plausible response. We select five clinical datasets (Healthfact, DDXPlus, SEER, BioMRC, Pubhealth) to illustrate the scaling law, and plot the results in Fig. 4. We observe that the first 4 samples provide significant

Select Model			Value Estimation				Traditional QA		Clinical Tasks			Mean
Majority	ORM	PRM	Min	Prod	Last	Vote-Sum	MedQA	MedMCQA	Healthfact	DDXPlus	SEER	
✓							70.93	64.21	70.37	57.00	52.19	62.94
	✓				✓		67.87	60.63	63.38	53.75	48.71	58.87
	✓					✓	71.88	64.67	70.12	57.15	53.44	63.45
		✓	✓				67.64	59.20	73.57	52.40	41.24	58.81
		✓		✓			65.51	60.00	75.51	57.30	47.67	61.20
		✓			✓		60.17	58.00	78.37	65.10	61.67	64.66
		✓				✓	**71.88**	**65.20**	**79.97**	**65.20**	**58.36**	**68.12**

Table 3: Comparison of P-VS with other decoding methods under the same token budgets. "Majority" means we use SC to select the final response. "ORM with Last" means selecting the response with the max ORM value.

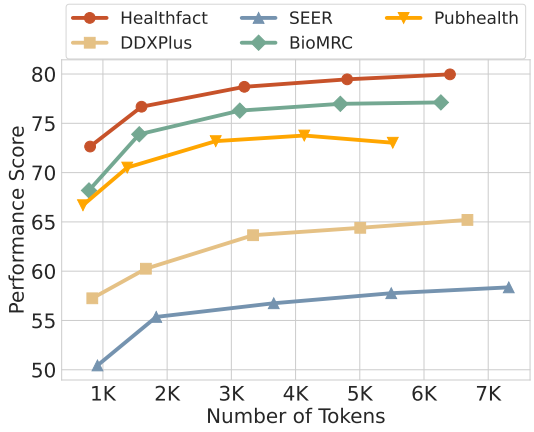

Figure 4: Budget scaling on clinical benchmarks. When the token budget increases, MedS[3]'s improves rapidly and then slows down as the budget is further increased.

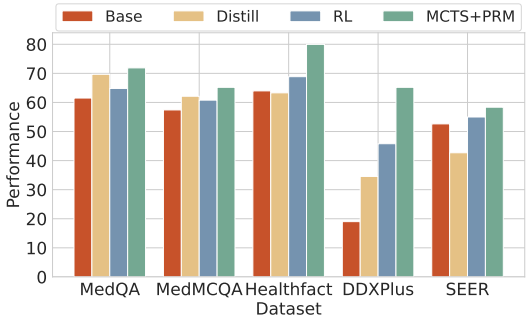

Figure 5: Three widely adopted methods to empower models with medical reasoning abilities. MCTS+PRM is the best among the three, making it the core of MedS[3].

performance improvements, showing a satisfactory tradeoff between the tokens consumption and the performance improvement. Although more generations bring fewer significant performance gains, the increasing trend never slows down except for the Pubhealth dataset, illustrating nearly unbounded scaling potentials.

5.3 Comparison of Slow-Thinking Styles

In this section, we compare MCTS plus PRM which is what MedS[3] leverages, with distillation from strong reasoning models, which is what O1-journey-part3 (Huang et al., 2025) does and pure reinforcement learning (RL), which is what DeepSeek-R1 (Guo et al., 2025) adopts. These methodologies are widely used for empowering small language models with strong long-chain reasoning abilities. We use the same dataset in §3 to implement RL, and use the officially released distillation dataset provided by Huang et al. (2025) to SFT the base model. The results presented in Fig. 5 demonstrate that in exam-level medical QA datasets where the base model already excels at, distillation from large proprietary reasoning models is much more data-efficient than the other two methods, albeit sacrificing generalization in clinical tasks. In contrast, with both considerable performance leap and generalization, RL is second to MCTS+PRM. We hypothesize that the medical diagnosis step is easier to determine than math reasoning steps, resulting in a more accurate PRM.

6 Conclusion

In this paper, we present MedS[3], a self-evolved slow-thinking medical language model built for universal clinical usage. We collect a seed dataset, with only 7,465 instances but covering 16 different realistic medical tasks, and use Monte-Carlo Tree Search to construct policy data and PRM data. We propose a new decoding method, which enables the resulting policy model to collaborate with the fine-tuned PRM model, to produce credible long-chain responses. Experiment results demonstrate that our model achieves superior performance on eleven downstream medical benchmarks, especially in realistic clinical ones, surpassing open-sourced models with a large margin.

Limitations

MedS^3 achieves superior performance over eleven benchmarks by conducting MCTS in seed datasets to collect both policy and PRM training data and a newly proposed decoding strategy: PRM-guided Vote-Sum. However, it can be further improved via these strategies: (1) conduct an iterative self-evolution pipeline to enhance both the policy and PRM; (2) cooperate with reinforcement learning to empower the policy with the "aha-moment" (Guo et al., 2025) ability; (3) introduce more training samples to avoid data imbalance.

Ethics Considerations

In developing clinical slow-thinking model MedS^3, it is crucial to address ethical consideration that arise when utilizing AI in healthcare environments. Below are the key ethical considerations that have been taken into account:

Perfomance vs. Potential Risks While MedS^3 demonstrates significant enhancements in clinical reasoning and task performance, it is important to acknowledge the inherent limitations of AI models. These models can generate misleading information of "hallucinations", which could pose risks in clinical settings. Therefore, MedS^3 is not intended to replace medical professionals or provide definitive clinical decisions but rather to assist healthcare providers under appropriate supervision.

Data Ethics and Privacy Compliance All patient data has been anonymized, and informed consent was obtained for its use, ensuring full compliance with privacy policies and obtaining explicit permission for all data usage. Additionally, data usage has been approved by relevant ethics committees, ensuring compliance with ethical standards and privacy protection requirements.

References

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

Anjanava Biswas and Wrick Talukdar. 2024. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint arXiv:2405.18346*.

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*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

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

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. 2023c. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock,

Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Snigdha Dubey, Gaurav Tiwari, Sneha Singh, Saveli Goldberg, and Eugene Pinsky. 2023. Using machine learning for healthcare treatment planning. *Frontiers in Artificial Intelligence*, 6:1124182.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

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

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Ziaofan Zhang. 2025. O1 replication journey – part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse drug reaction classification with deep neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 877–887, Osaka, Japan. The COLING 2016 Organizing Committee.

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

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

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

HyoJe Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, et al. 2024. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *arXiv preprint arXiv:2404.05144*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

DKI Lab. 2020. covid19-classification: Document classification on covid-19 literature using the litcovid collection and the hedwig library. `https://github.com/dki-lab/covid19-classification`.

Yusheng Liao, Shuyang Jiang, Zhe Chen, Yu Wang, and Yanfeng Wang. 2024a. MedCare: Advancing medical LLMs through decoupling clinical alignment and knowledge aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10562–10581, Miami, Florida, USA. Association for Computational Linguistics.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024b. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

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*.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, page ocae037.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. 2025. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*.

Clara H McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: matching user questions to covid-19 faqs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3458–3465.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. 2024. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Website. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Learning to reason with large language models. Accessed: 2024-01-15.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

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

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus Dataset.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. 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 V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Mark HM Winands, Yngvi Björnsson, and Jahn-Takeshi Saito. 2008. Monte-carlo tree search solver. In *Computers and Games: 6th International Conference, CG 2008, Beijing, China, September 29-October 1, 2008. Proceedings 6*, pages 25–36. Springer.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024b. Towards evaluating and building versatile large language models for medicine. *arXiv preprint arXiv:2408.12547*.

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024c. Towards evaluating and building versatile large language models for medicine. *Preprint*, arXiv:2408.12547.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024d. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

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

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, Jin Lu, Wei Zhang, Tuo Zhang, Lu Zhang, Dajiang Zhu, Xiang Li, Wei Liu, Quanzheng Li, Andrea Sikora, Xiaoming Zhai, Zhen Xiang, and Tianming Liu. 2024a. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *CoRR*, abs/2411.14461.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. 2025. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *arXiv preprint arXiv:2501.09213*.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*.

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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

A Related Works

With the success of the generalist LLMs, their advancements in both clinical and biomedical scenarios have shown significant promise. Numerous previous works focus on developing medical-specific LLMs, which are now being increasingly adopted across various clinical settings. These medical LLMs generally follow three main approaches:

Continual Pre-training Medical LLMs These types of medical LLMs (Xu, 2023; Chen et al., 2023c) are developed on the advantage of generalist

LLM and attempt to inject domain-specific knowledge and expertise through continual pre-training techniques. Such type of methods usually require significant computational resources. For example, 3 billion training tokens are used in HuatuoGPT-II (Chen et al., 2023a) and PMC-Llama (Wu et al., 2024a) even requires more than 75 billion tokens. However, results in recent works (Qiu et al., 2024) show that the benefits of continued pre-training are diminishing as the capabilities of the generalist LLMs improve.

Fine-tuned Medical LLMs Compared to continuous pre-training, fine-tuning is a more efficient approach. It can rapidly adapt to medical scenarios and perform the relevant tasks effectively when the base LLMs are sufficiently powerful. (Ankit Pal, 2024; Christophe et al., 2024; Zhang et al., 2024b) Specifically, Liao et al. (2024a) develops a two-stage method which can decouple the knowledge-injection and clinical alignment procedure during the fine-tuning process to prevent the 'alignment-tax.' Wu et al. (2024c) collects a wide range of medical language processing tasks spanning 19 task categories and 122 unique clinical tasks to improve the LLMs' capacities on various downstream clinical tasks.

Slow-Thinking Medical LLMs With the significant achievements of the o1 (Jaech et al., 2024) in complex reasoning tasks, previous works show the potential advantage of the o1-like models in medical tasks (Xie et al., 2024; Xu et al., 2024a; Nori et al., 2024). Based on these, previous works develop the slow-thinking medical LLMs with distillation: Huang et al. (2025) directly learn the reasoning trajectory generated by o1 and Chen et al. (2024) improving the model's reasoning ability through o1 synthesis of reflective data and reinforcement learning. Besides, Yu et al. (2025) create a Chinese version slow-thinking medical LLMs by constructing the preference data with QwQ (Team, 2024).

B Further Experiments

In this section, we present more experiments to validate the effectiveness of MedS^3.

B.1 Soft Train vs. Hard Train

We here compare the PRM training techniques, where the step label is set to a hard two-class $0/1$ label (Zhang et al., 2025) or a soft float label introduced in MedS^3. We remain the other settings unchanged and set the label to 1 if the step's rollout value is greater than 0, otherwise 0 when training the hard PRM. The results in Table 4 indicate that soft-label is comprehensively superior to hard-label. Although Zhang et al. (2025) indicate that hard-label works after prompting an LLM to filter bad intermediate steps, for the sake of privacy in clinical uasge and fairness we do not conduct filtering.

Method	MedQA	MedMCQA	Healthfact	DDXPlus	SEER	Average
Hard	64.81	62.18	77.61	64.35	54.94	64.778
Soft	71.88	65.20	79.97	65.20	58.36	68.122

Table 4: Comparison between soft-label which is adopted in MedS^3 and hard-label.

B.2 Reliability of PRM

Although our process reward model (PRM) is trained using rollout-value—a rule-based supervision signal—we empirically demonstrate that the PRM exhibits strong capability in identifying erroneous reasoning steps. To evaluate this, we compare two metrics: the task error rate (defined as instances where model predictions deviate from ground truth answers) and the PRM error rate (occurring when the PRM assigns a higher score to an incorrect candidate than to the ground truth-aligned prediction). As illustrated in Table 5, the majority of tasks exhibit a PRM error rate of zero, with the exception of the Healthfact dataset. This suggests that our fine-tuned PRM effectively differentiates between valid and invalid clinical reasoning steps. Furthermore, the observed discrepancy between task and PRM error rates implies that most errors arise not from the PRM's assessment but from the policy model's failure to generate candidates aligned with the ground truth. This finding underscores the need to refine the policy model in future iterations to address this limitation.

C Future Work

As a pioneering work, we have validated that small language models can self-evolve to empower themselves with strong reasoning abilities in clinical usage. There are several remaining directions to further enhance MedS^3:

1. Conduct iterative evolution. Currently we are only focusing on one iteration evolution, which greatly leaves the PRM under-tuned.

2. Conduct Human-interference evaluation. MC-

Error Type	MedQA	MedMCQA	HealthFact	PubmedQA	BioASQ	Med MMLU	BioMRC	DDXPlus	PubHealth	DrugDose	SEER
Task	28.91	34.60	20.03	40.40	23.79	20.50	22.88	34.60	26.97	4.35	41.87
PRM	0.00	0.00	16.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: Comparison of error rates (lower is better) in task-level and PRM-level, where PRM works with no errors in most of testbeds.

rollout value is verified to be not the best choice for evaluating the value of an internal step. We are eager to introduce fine-grained step label to enhance the optimization of the PRM.

3. Introduce more clinical data, not limited to close-ended generation. Currently, all the data used in MedS^3 are close-ended, and the application of reasoning is not limited to such a narrow room. We intend to extend MedS^3 to broader clinical tasks to make MedS^3 a more useful system.

We will continue our exploration and make MedS^3 more practical in medical domains.

D Prompt Template

We show the prompt used to synthesize reasoning data in Fig. 6 and Fig. 7.

E Dataset Details

In this section, we elucidate the seed dataset and the evaluation sets. We divide the used 16 training datasets into the following five dimensions:

1. **Long Context QA**: This dimension enables MedS^3 to capture useful information from the given context and response with long-chain reasoning. This dimension covers BioMRC (Pappas et al., 2020), HeadQA Topic Classification (Vilares and Gómez-Rodríguez, 2019; Wu et al., 2024b), and DDXPlus (Tchango et al., 2022).

2. **Knowledge-Intensive QA**: This dimension teaches MedS^3 to use long-chain reasoning to answer knowledge-intensive problems, which covers MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019) and PubHealth (Kotonya and Toni, 2020).

3. **Bio-Medical QA**: This part leverages general data in bio-medicine domains to enhance the generality of MedS^3, which includes SciQ (Welbl et al., 2017), Evidence Inference (DeYoung et al., 2020) and Head QA (Vilares and Gómez-Rodríguez, 2019).

4. **Medical Natural Language Inference**: This dimension prompts MedS^3 to discriminate biomedical research concepts and corresponding descriptions, which contains Healthfact Classification (Kotonya and Toni, 2020), Medical Question Pair (MQP; McCreery et al. (2020)), and catalonia-independence-corpus (CIC; Zotova et al. (2020)).

5. **Diagnosis QA**: This dimension is related to real-world clinical scenarios, including disease diagnosis and classification and drug related questions. We choose Covid-19 Classification (Lab, 2020), Drug-Dose Extraction and Adverse Drug Event Classification (Huynh et al., 2016; Wu et al., 2024b).

The descriptions of each training and evaluation datasets are presented below:

1. MedQA (Jin et al., 2021) is a widely used benchmark for evaluating AI systems in medical question answering, featuring multiple-choice questions from professional medical licensing exams such as the USMLE and exams from China and Taiwan. We adopt its 5-options English version, taking its training set as seed data and 1,273 test problems as the evaluation benchmark.

2. PubmedQA (Jin et al., 2019) is a specialized benchmark for biomedical question answering, consisting of question-answer pairs derived from PubMed abstracts. It focuses on yes/no/maybe questions that require reasoning over biomedical literature. We use the human-labeled question set and split the training set and test set, with both 500 problems for evolution and evaluation, respectively. Note that we do not include relevant contexts before questions, challenging models' internal knowledge comprehension.

3. MedMCQA (Pal et al., 2022) is a large-scale benchmark for medical question answering,

Reason Template

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023 Today Date: 23 July 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Reasoning Example: {Few-shot Example}

You are a professional medical expert majored at reasoning in hard
medical-related problems.

Think critically about the problem and answer with concise, accurate reasoning.
Please ensure your reasoning is thorough and elaborate, breaking down each step
of your thought process.

Problem: {problem}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Step 0: Let's break down this problem step by step

Step 1:
```

Figure 6: Reason template

Finish Template

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023 Today Date: 23 July 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Reasoning Example: {Few-shot Example}

You are a professional medical expert majored at reasoning in hard
medical-related problems.

Use thorough and elaborate steps to complete your reasoning. Conclude the task
by stating: "The answer is {answer}".

Problem: {problem}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Step 0: Let's break down this problem step by step

Step 1:
```

Figure 7: Finish template

featuring over 194,000 multiple-choice questions sourced from Indian medical entrance exams and other educational resources. It spans a wide range of medical topics, including anatomy, pharmacology, and pathology, and is designed to evaluate the reasoning and knowledge application skills of AI systems in a clinical context. The test set contains 4,183 problems.

4. MMLU (Hendrycks et al., 2021) is to measure LLM's multitask accuracy, which contains 14,421 problems. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more. We select its medical-related problems, resulting in a test set with 1,561 problems.

5. BioMRC (Pappas et al., 2020) is a collection of medical-related question-answer pairs, specifically designed for the evaluation of machine reading comprehension (MRC) tasks in the biomedical domain. It is derived from a wide range of medical texts, including clinical notes, research papers, and medical textbooks. The dataset contains a series of questions and corresponding answers, where the answers are extracted from relevant passages of text. We use its 6,250 test set as the evaluation set.

6. HeadQA (Vilares and Gómez-Rodríguez, 2019) is a specialized medical question-answering dataset designed to evaluate models in the context of neurology and head-related disorders. It consists of a collection of questions paired with answers derived from a variety of clinical notes, medical reports, and other head-related health data sources.

7. DDX-Plus (Tchango et al., 2022) is a comprehensive medical diagnostic dataset designed to assist in the development and evaluation of machine learning models for differential diagnosis in clinical settings. It consists of clinical cases, where each case includes a set of symptoms, patient history, physical examination findings, and diagnostic questions, along with a list of potential diagnoses ranked by their likelihood. The diverse set of cases in the dataset spans multiple medical specialties, making it an ideal resource for creating models capable of assisting healthcare professionals in making informed diagnostic decisions. Due to its huge test set (over 100,000 test instances), we randomly select 2,000 items for evaluation.

8. SciQ (Welbl et al., 2017) is a scientific question-answering dataset designed to assess machine learning models in answering factual questions across a wide range of scientific domains. It consists of over 13,000 questions derived from scientific literature, including topics in physics, biology, chemistry, and earth sciences, among others. Each question is paired with a correct answer and is supported by a passage of text from which the answer is extracted.

9. Evidence Inference (DeYoung et al., 2020) is a collection designed to evaluate machine learning models on their ability to infer logical conclusions from evidence presented in the form of textual information. This dataset consists of structured pairs of premises (evidence) and hypotheses, where the goal is for models to determine the logical relationship between them—whether the hypothesis is supported, contradicted, or is neutral with respect to the provided evidence. Typically used for tasks such as textual entailment or natural language inference (NLI), the dataset includes a variety of complex scenarios across multiple domains, including law, healthcare, and science, where reasoning based on available evidence is crucial.

10. PubHealth (Kotonya and Toni, 2020) is a comprehensive dataset for explainable automated fact-checking of public health claims. Each instance in the PUBHEALTH dataset has an associated veracity label (true, false, unproven, mixture). Furthermore each instance in the dataset has an explanation text field. The explanation is a justification for which the claim has been assigned a particular veracity label. We construct two different test sets. Healthfact is to directly predict whether a given instance is true/false/unproven/mixture. The other, Pubhealth, is to predict whether the instance sentence and the given explanation expresses the same meaning.

11. Medical Question Pair (McCreery et al., 2020) contains a dataset of 3,048 similar and dissimilar medical question pairs hand-generated and

labeled by Curai's doctors. Models should clarify whether the given two questions are similar or not.

12. Catalonia-independence-Corpus (Zotova et al., 2020) is a dataset built for stance detection in Twitter for the Catalan and Spanish languages, with the aim of facilitating research on stance detection in multilingual and cross-lingual settings.

13. Covid-19 Classification (Lab, 2020) is an extension of the Hedwig library and contains all necessary code to reproduce the results of some document classification models on a COVID-19 dataset created from the LitCovid collection.

14. Adverse Drug Event (Huynh et al., 2016) is critical for developing automated systems that can support clinicians in identifying harmful drug reactions, potentially reducing healthcare costs, and enhancing patient safety. Given the increasing volume of clinical data, this dataset plays a key role in advancing AI-driven drug safety research and improving the overall quality of healthcare. We build Drugdose extraction test set to benchmark models to extract the exact dose of a specific drug.

F Hyperparameters

F.1 Data Synthesis

For each node expansion, we simultaneously generate 3 different responses with the same prompt. We set the generation temperature to 1. The stop tokens are set to $\{\texttt{Step k:} \mid k = 1, 2, \cdots 100\}$ to ensure that each node represents a single reasoning step. We use the first sample in MedQA as the one-shot example and prompt GPT-4o to generate step-by-step outputs.

F.2 Self-Training of Policy and PRM

Policy tuning We use trl[3] as the training framework. We use vanilla `SFTTrainer` to train the policy model. We set the warmup ratio to 0.03 and the max sequence length to 8192. The batch size is set to 64 and the learning rate is set to 2e-4. We use LoRA (Hu et al., 2022) to efficiently train the model, where the rank r and alpha α are set to 16 and 32, respectively. We train 1 epoch for the total 24,441 positive synthesized samples.

[3]https://huggingface.co/docs/trl/index

PRM tuning We use `PRMTrainer` of trl to train the PRM model. We use the same hyperparameters as those used in the policy fine-tuning but conduct a simple data filtering method. For an instance where all its solution trajectories are all correct or incorrect, we only randomly maintain one of them, as the excessively simple or hard instance will disturb the training process. For other instances, we guarantee that the correct trajectories and incorrect trajectories have the same total count for a given data point, which performs a simple 0-1 class balance. We train 1 epoch for filtered 76,792 samples with LoRA, where the rank r and α are set to 64 and 128, respectively.

F.3 Evaluation

For evaluation, the temperature is set to 1.0 and top_p is set to 0.9. The max generation tokens are set to 8,192. Among the three presented decoding mechanisms, CoT (Wei et al., 2022) directly prompts models to generate a long reasoning chain and outputs the answer with "The answer is {answer}" for the convenience of answer extraction. Self-Consistency (Wang et al., 2023) generates $N = 16$ samples for a given problem, and we select the one whose answer appears most times among the N outputs. For PRM-guided Vote-Sum, we generate $N = 12$ samples for datasets where SC has already achieved a high score (MedMCQA, Med-MMLU and BioASQ) and generate $N = 16$ samples for other datasets. This is reasonable since not all generated samples are of high quality due to the sampling parameters `top_p`. We manually filter some low-probability samples to prevent them from disturbing the Vote-Sum computation. We use exact match (EM) to measure the performance. Specifically, we extract the contents following the last "The answer is" template to match the self-reflection thinking style, and perform appropriate post-process to derive a final prediction. For multiple-choice problems, we directly choose the first character of prediction phrases and measure whether the ground truth is equal to the prediction. For close-ended generation tasks, we remove quotes and turn the prediction and the ground truth into lowercase phrases. After that, we check whether the ground truth phrases exist in the prediction phrases.

F.4 Training Details of Distillation and RL

In this section, we elucidate the implementation details of distillation and RL.

Distillation For Distillation method, we fine-tune Llama3.1-8B with 2K training data[4] released by Huang et al. (2025), which combined with the questions in MedQA and corresponding response generated by o1 (Jaech et al., 2024). We adopt LoRA (Hu et al., 2022) and set the rank r to 16 and alpha α to 32 for fair comparisons. For fine-tuning parameters, we set the learning rate to $2e$-6 and batch size to 128.

RL We follow Guo et al. (2025) to use Group Relative Policy Optimization (GRPO; Shao et al. (2024)) to conduct RL training. We set the number of generations to 10 and the learning rate to $1e-6$. We adopt ZeRO-3 (Rajbhandari et al., 2020) to save memory and conduct full fine-tuning in one 8xA100 machine. The batch size is set to 4 per GPU. For the adopted prompt, we use the same prompt illustrated in DeepSeek-R1-zero, and use `<think></think><answer></answer>` to learn the slow-thinking output style. We use accuracy reward and format reward, and the reward setting is presented below in Table 6.

Table 6: Reward configurations for training RL-style medical models. Null means that there is no contents between `<answer>` and `</answer>`.

Conditions	Accuracy	Format
Correct	1	1
Incorrect	-1	-1
Null	0	Null

G AI Assistant Usage

We mainly use AI assistants to polish our expressions.

H Case Study

[4] https://huggingface.co/datasets/SPIRAL-MED/o1-journey-Ophiuchus

Question:
A 47-year-old executive schedules an appointment his physician for a routine medical check-up. He currently has no complaints and claims to be “as fit as a fiddle.” The physical examination findings are unremarkable, except for a mid-systolic murmur heard in the 2nd left intercostal space that radiates to the carotids on auscultation. The physician instructs the patient to stand from a supine position with the stethoscope still placed on his chest. Which of the following changes would occur with this maneuver?\nA. An increase in right atrial pressure\nB. An increase in left ventricular end-diastolic pressure\nC. An increase in pulmonary capillary wedge pressure\nD. A reduction in the slope of the pacemaker potential\nE. A reduction in diastolic filling time

Llama3.1-8B:
When the patient stands from a supine position, the following changes occur:\n\n* Venous return to the heart decreases due to gravity, which leads to a decrease in right atrial pressure (A).\n* The decrease in venous return also leads to a decrease in left ventricular end-diastolic pressure (B).\n* The decrease in left ventricular end-diastolic pressure would not directly affect the pulmonary capillary wedge pressure (C).\n* The pacemaker potential is not directly affected by changes in posture.\n* The diastolic filling time is not directly affected by changes in posture.\n\nThe mid-systolic murmur heard in the 2nd left intercostal space that radiates to the carotids is likely due to aortic stenosis. When the patient stands from a supine position, the murmur may become louder due to the increased gradient across the aortic valve. \nThe answer is A.

MedS[3]
Step 0: Let's break down this problem step by step.\n\nStep 1: Identify the likely diagnosis based on the physical examination findings. The presence of a mid-systolic murmur heard in the 2nd left intercostal space that radiates to the carotids on auscultation is suggestive of aortic stenosis. Aortic stenosis is characterized by the narrowing of the aortic valve opening, which can lead to a delay in left ventricular ejection and a subsequent murmur.\n\nStep 2: Understand the effect of standing from a supine position on the cardiovascular system. When a person stands from a supine position, gravity causes a decrease in venous return to the heart. This decrease in venous return leads to a decrease in the preload, which is the initial stretching of the cardiac myocytes prior to contraction.\n\nStep 3: Determine how the decrease in preload affects the heart's function. A decrease in preload leads to a decrease in the end-diastolic volume, which in turn leads to a decrease in the stroke volume. This is because the heart's ability to pump blood is directly related to the amount of blood it receives during diastole.\n\nStep 4: Consider how the decrease in stroke volume affects the murmur associated with aortic stenosis. The murmur associated with aortic stenosis is typically a systolic ejection murmur, which is caused by the delay in left ventricular ejection. When the stroke volume decreases, the murmur may become softer or less intense, but it will not change its timing or location.\n\nStep 5: Analyze the options provided to determine which one is most closely related to the change in the murmur associated with aortic stenosis. The murmur is not directly related to the options provided, but the decrease in stroke volume is related to the decrease in diastolic filling time. When the heart is in a supine position, the diastolic filling time is longer due to the increased venous return. When the person stands, the diastolic filling time decreases due to the decreased venous return.\n\nStep 6: Select the correct answer based on the analysis. The correct answer is the one that is most closely related to the change in the murmur associated with aortic stenosis, which is the decrease in diastolic filling time.\n\n**The answer is E.**

MedS[3]-PRM:
Step 0: Let's break down this problem step by step.\n\nStep 1: Identify the likely diagnosis. The presence of a mid-systolic murmur heard in the 2nd left intercostal space that radiates to the carotids suggests aortic stenosis. This condition is characterized by the narrowing of the aortic valve opening, leading to obstruction of blood flow from the left ventricle into the aorta.\n\nStep 2: Understand the impact of changing the patient's position on the heart's hemodynamics. The physician instructs the patient to stand from a supine position. This change in position affects the heart's position and the volume of blood returning to the heart. When a person stands up, gravity causes blood to pool in the lower extremities, reducing the volume of blood returning to the heart.\n\nStep 3: Analyze the effect on venous return to the heart. Standing from a supine position reduces the volume of blood returning to the heart due to venous pooling in the lower extremities. This reduction in venous return affects the heart's ability to fill with blood, particularly during diastole.\n\nStep 4: Relate the change in venous return to the heart's hemodynamics. The reduced venous return to the heart results in decreased diastolic filling time. This is because the heart has less blood available to fill with during diastole, leading to a shorter diastolic filling period.\n\nStep 5: Eliminate incorrect options. The other options do not directly relate to the change in position and its effect on the heart's hemodynamics. Option A (an increase in right atrial pressure) might occur in conditions of volume overload or tricuspid regurgitation but is not directly related to the change in position. Option B (an increase in left ventricular end-diastolic pressure) and Option C (an increase in pulmonary capillary wedge pressure) might occur in conditions of left ventricular failure or pulmonary congestion but are not directly related to the change in position. Option D (a reduction in the slope of the pacemaker potential) is not directly related to the change in position and is more related to electrophysiological changes.\n\nStep 6: Choose the correct answer. The correct answer is related to the change in diastolic filling time due to reduced venous return to the heart.\n\n**The answer is E.**

Figure 8: Cases of MedQA

Question:
Imagine you are a doctor, next I will give you a dialogue and please make a diagnosis decision based on it. You only need to choose one answer among the following lists: ['spontaneous pneumothorax', 'cluster headache', 'boerhaave', 'spontaneous rib fracture' 'gerd', 'hiv (initial infection)', 'anemia', 'viral pharyngitis', 'inguinal, hernia', 'myasthenia gravis', 'whooping cough', 'anaphylaxis', 'epiglottitis', 'guillain-barré syndrome', 'acute laryngitis', 'croup', 'psvt', 'atrial fibrillation', 'bronchiectasis', 'allergic sinusitis', 'chagas', 'scombroid food poisoning', 'myocarditis', 'larygospasm', 'acute dystonic reactions', 'localized edema', 'sle', 'tuberculosis', 'unstable angina', 'stable angina', 'ebola', 'acute otitis media', 'panic attack', 'bronchospasm / acute asthma exacerbation', 'bronchitis', 'acute copd exacerbation / infection', 'pulmonary embolism', 'urti', 'influenza', 'pneumonia', 'acute rhinosinusitis', 'chronic rhinosinusitis', 'bronchiolitis', 'pulmonary neoplasm', 'possible nstemi / stemi', 'sarcoidosis', 'pancreatic neoplasm', 'acute pulmonary edema', 'pericarditis', 'cannot decide']. The answer should be a single word. \nAge: 51 \n\nSex: M \n\nInitial evidence: Have you been coughing up blood? Yes \n\nEvidence: Have you been coughing up blood? Yes;Do you have pain somewhere, related to your reason for consulting? Yes;Characterize your pain: Sensitive;Characterize your pain: A knife stroke;Do you feel pain somewhere? Posterior chest wall(r);Do you feel pain somewhere? Posterior chest wall(l);How intense is the pain? 5;Does the pain radiate to another location? Nowhere;How precisely is the pain located? 4;How fast did the pain appear? 5;Are you experiencing shortness of breath or difficulty breathing in a significant way? Yes;Do you smoke cigarettes? Yes;Do you constantly feel fatigued or do you have non-restful sleep? Yes;Have you recently had a loss of appetite or do you get full more quickly then usually? Yes;Have you had an involuntary weight loss over the last 3 months? Yes;Are you a former smoker? Yes;Do you have a cough? Yes;Have you traveled out of the country in the last 4 weeks? N;Are you exposed to secondhand cigarette smoke on a daily basis? Yes;Do you have family members who have had lung cancer? Yes.

Llama3.1-8B:
Based on the provided dialogue, I would choose: 'bronchogenic carcinoma'. \nThe answer is 'bronchogenic carcinoma' because the patient is a 51-year-old male who has been coughing up blood, has a history of smoking, is a former smoker, and has a family history of lung cancer.

MedS^3
Step 0: Let's break down this problem step by step.\n\nStep 1: The patient presents with coughing up blood (hemoptysis), which is a symptom that can be associated with several conditions, including lung cancer, pulmonary embolism, and infections such as pneumonia or tuberculosis. The patient's age (51) and smoking history (current and former smoker) increase the likelihood of lung cancer.\n\nStep 2: Assess the patient's pain characteristics. The patient reports pain in the posterior chest wall, which is a common location for lung cancer. The pain is described as sensitive and like a knife stroke, which suggests that it may be related to a tumor or mass in the lung.\n\nStep 3: Evaluate the patient's other symptoms, including shortness of breath, fatigue, loss of appetite, and weight loss. These symptoms are consistent with a diagnosis of lung cancer, particularly if the patient has a smoking history.\n\nStep 4: Consider the patient's family history of lung cancer, which may increase the risk of developing the disease.\n\nStep 5: Based on the patient's symptoms and risk factors, the most likely diagnosis is lung cancer. The patient's cough and exposure to secondhand cigarette smoke also support this diagnosis.\n\nStep 6: Among the provided options, the most likely diagnosis is 'pulmonary neoplasm', which is a broad term that includes lung cancer.\n\nThe answer is pulmonary neoplasm. \n**The answer is pulmonary neoplasm.**

MedS^3-PRM:
Step 0: Let's break down this problem step by step.\n\nStep 1: The patient is a 51-year-old male with a history of smoking, exposure to secondhand cigarette smoke, and a family history of lung cancer. He presents with symptoms of coughing up blood, chest pain, shortness of breath, fatigue, loss of appetite, and involuntary weight loss. These symptoms suggest a serious underlying condition that requires immediate attention.\n\nStep 2: The patient's symptoms of coughing up blood and chest pain are indicative of a respiratory or cardiac issue. Given his history of smoking and exposure to secondhand smoke, lung cancer is a significant concern. However, other conditions such as chronic obstructive pulmonary disease (COPD), pneumonia, or cardiac conditions like myocardial infarction or pericarditis should also be considered.\n\nStep 3: The patient's chest pain is described as sensitive and like a knife stroke, localized to the posterior chest wall on both sides. This pain is severe, appearing rapidly, and is associated with shortness of breath. These symptoms suggest a condition that is causing inflammation or obstruction of the airways or lungs.\n\nStep 4: Considering the patient's symptoms and history, the most likely diagnosis is lung cancer. However, other conditions such as COPD, pneumonia, or cardiac conditions should also be considered. Given the patient's coughing up blood, chest pain, and shortness of breath, a diagnosis of lung cancer is the most likely explanation for his symptoms.\n\nStep 5: Based on the patient's symptoms and history, the best answer is 'pulmonary neoplasm'. This diagnosis takes into account the patient's coughing up blood, chest pain, shortness of breath, and history of smoking and exposure to secondhand smoke.\n\n**The answer is 'pulmonary neoplasm'.**

Figure 9: Cases of DDXPlus

Question:
Imagine you are a doctor, next I will give you summarization of a patient and please help me to give a suggestion for next treatment. You just need to pick out an answer from ['Intraoperative rad with other rad before/after surgery', 'Intraoperative radiation', 'No radiation and/or cancer-directed surgery', 'Radiation after surgery', 'Radiation before and after surgery', 'Radiation prior to surgery', 'Surgery both before and after radiation']\nAge at diagnosis: 73\n\nRegional nodes positive (1988+): 0\n\nTotal number of in situ/malignant tumors for patient: 1\n\nRadiation recode: None/Unknown\n\nChemotherapy recode: No/Unknown\n\nER Status Recode Breast Cancer (1990+): Positive\n\nPR Status Recode Breast Cancer (1990+): Negative\n\nCS tumor size (2004-2015): 16\n\nDerived HER2 Recode (2010+): Negative\n\nRegional nodes examined (1988+): 8\n\nCOD to site recode: Diseases of Heart\n\nRace recode: White\n\nSex: Female\n\nVital status recode (study cutoff used): Dead\n\nDiagnosis_year: 2010\n\nLast_fu _year: 2015\n\ninterva_years: 5

Llama3.1-8B:
Based on the provided information, I would suggest the following treatment plan:\n\nGiven the patient's age (73) at diagnosis, the presence of a single tumor (1), and the fact that the patient has died from a cause other than breast cancer (Diseases of Heart), it's likely that the cancer was not the primary cause of death.\n\nHowever, considering the patient's ER-positive status and the relatively large tumor size (16), a treatment plan that includes radiation therapy may be beneficial. \n\nGiven the patient's age and the fact that the cancer was not the primary cause of death, I would suggest 'Radiation after surgery' as the next treatment plan. This approach would allow for the removal of the tumor through surgery, followed by radiation therapy to target any remaining cancer cells and reduce the risk of recurrence. \nThe answer is 'Radiation after surgery'..

MedS[3]
Step 0: Let's break down this problem step by step.\n\nStep 1: Identify the key information from the patient's summary. The patient is a 73-year-old female with breast cancer, diagnosed in 2010. The cancer is ER-positive, PR-negative, and HER2-negative. The patient had 8 regional nodes examined, but none were positive. The cancer was 16mm in size.\n\nStep 2: Assess the patient's prognosis and treatment options. Given the patient's age, cancer size, and ER-positive status, the patient is likely to benefit from hormone therapy. However, the patient's PR-negative status suggests that PR-targeted therapy may not be effective.\n\nStep 3: Consider the patient's vital status and last follow-up year. The patient passed away in 2015, which suggests that the cancer may have progressed despite treatment.\n\nStep 4: Evaluate the treatment options for ER-positive breast cancer. The standard treatment for ER-positive breast cancer includes hormone therapy, such as tamoxifen or aromatase inhibitors. However, the patient's PR-negative status and cancer progression suggest that hormone therapy may not be sufficient.\n\nStep 5: Consider the patient's cancer-directed surgery and radiation therapy. The patient's cancer size and regional nodes examined suggest that surgery and radiation therapy may have been used to treat the cancer.\n\nStep 6: Evaluate the radiation therapy options. Given the patient's cancer progression, radiation therapy may have been used to control symptoms or slow disease progression.\n\nStep 7: Consider the patient's overall health and treatment goals. The patient's age and cancer progression suggest that aggressive treatment may not be feasible.\n\nStep 8: Based on the patient's cancer characteristics and treatment history, the most likely treatment approach is radiation after surgery.\n\nThe answer is 'Radiation after surgery'.

MedS[3]-PRM:
Step 0: Let's break down this problem step by step.\n\nStep 1: Analyze the patient's information to identify relevant details that may impact treatment decisions. The patient is a 73-year-old female with a diagnosis of breast cancer in 2010, which is now 5 years after last follow-up. Key information includes:\n\n- ER-positive, PR-negative breast cancer, indicating hormone receptor status.\n- HER2-negative, which may impact targeted therapy options.\n- 1 in situ/malignant tumor, suggesting a localized cancer.\n- No regional nodes positive, indicating no distant metastasis.\n- The patient's vital status is recorded as dead, but the cause of death is listed as diseases of the heart, which may not be directly related to breast cancer.\n\nStep 2: Determine the patient's cancer stage and potential treatment options. Given the patient's ER-positive status, treatment options may include hormone therapy, but the lack of information on chemotherapy recode and radiation recode makes it challenging to determine the extent of treatment.\n\nStep 3: Consider the patient's age, comorbidities, and overall health status, which may influence treatment decisions. The patient is 73 years old, and the cause of death is listed as heart disease, suggesting potential comorbidities.\n\nStep 4: Evaluate the impact of HER2-negative status on treatment options. HER2-negative breast cancers are typically treated with hormone therapy, chemotherapy, or a combination of both.\n\nStep 5: Consider the role of surgery in the patient's treatment plan. While the specific details of the patient's cancer-directed surgery are not provided, it is likely that surgery played a role in the patient's treatment.\n\nStep 6: Assess the potential benefits and risks of radiation therapy in this patient. Given the patient's age and comorbidities, the decision to use radiation therapy may be influenced by the potential benefits and risks.\n\nStep 7: Determine the most appropriate next treatment step based on the patient's information. Considering the patient's ER-positive status, the lack of information on chemotherapy recode and radiation recode, and the patient's age and comorbidities, it is challenging to determine the most appropriate next treatment step.\n\nHowever, given the lack of information on radiation therapy and chemotherapy, and considering the patient's age and comorbidities, I would suggest the following treatment option:\n\nGiven the patient's ER-positive status and the lack of information on chemotherapy recode and radiation recode, I would recommend a treatment option that focuses on hormone therapy and cancer-directed surgery. However, without more information on the patient's cancer stage and treatment history, it is difficult to provide a more specific recommendation.\n\nAmong the provided options, the most conservative approach would be to choose the option that does not involve radiation therapy, considering the patient's age and comorbidities.\n\n**The answer is 'No radiation and/or cancer-directed surgery'**

Figure 10: Cases of SEER