MedS^3: Towards Medical Slow Thinking with Self-Evolved Soft Dual-sided Process Supervision

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 o-series models highlight 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 small-scale medical reasoning system, MedS^3, designed for long-chain reasoning in clinical tasks using a self-evolution paradigm. Starting with 8,000 instances sampled with a curriculum strategy 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 for two iterations. Each reasoning step is scored by the rollout estimation, allowing for training the policy model and a soft dual-sided process reward model (PRM). Experiments on eleven evaluation datasets demonstrate that MedS^3 outperforms not only the prior strongest medical model by 6.45, but also 32B-level general reasoning models by 8.57 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), supporting tasks from clinical note generation (Biswas and Talukdar, 2024; Jung et al., 2024) to patient communication (Tu et al., 2024; Liao et al., 2024b). Despite these advances, enabling reliable and robust long-chain reasoning remains a critical challenge for medical-oriented language models (MLMs), which is essential for clinical decision-making where each reasoning step must be accurate, interpretable, and evidence-based.

Recent progress in "slow-thinking" reasoning models such as OpenAI's o-series (OpenAI, 2024) has led to breakthroughs on complex reasoning tasks (Lyu et al., 2025; Wang et al., 2024). However, these advances have not yet been fully realized in medical language models, where domain complexity, data scarcity, and the need for rigorous stepwise justification amplify the difficulty. Prior efforts to improve MLMs have focused heavily on large-scale pretraining (Qiu et al., 2024), which requires enormous computational resources but yields only modest task improvements. Alternatively, supervised fine-tuning (SFT) on human-annotated datasets (Ouyang et al., 2022) often involves concise, single-answer outputs that can harm language fluency and limit the model's ability to generate rich, stepwise clinical reasoning necessary for real-world applications.

Synthetic datasets generated by LLMs offer a promising direction but suffer from hallucinations (Xu et al., 2024b; Huang et al., 2023), which constrain the effective learning of clinically sound reasoning paths. 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 heavy reliance on large proprietary models (GPT-series) limits its generalizability to real clinical applications for the sake of data privacy. 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. These challenges

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.

highlight a core problem: how to efficiently induce robust, interpretable, and stepwise reasoning capabilities in small-scale medical models without relying on prohibitive pretraining, proprietary models or noisy synthetic supervision.

To address this, we propose MedS^3, a novel small-scale medical language model that integrates robust long-chain reasoning "policy" with a fine-grained, soft dual-sided Process Reward Model (PRM) designed to evaluate and guide reasoning steps progressively. We first curate 16 medical tasks spanning clinical diagnosis QA, natural language inference, knowledge-intensive QA, long-context QA, and biomedical QA, and design a curriculum sampling strategy to evolve the model over 8,000 carefully selected challenging instances per iteration. Upon these, our key innovation involves a self-bootstrapping pipeline leveraging a Monte-Carlo Tree Search (MCTS)-based reflection-aware evolution process (§2.1) that iteratively generates high-quality synthetic data for both policy fine-tuning (§2.2) and soft dual-sided PRM training. To enable reflection-aware step-wise supervision, we design a soft dual-sided label (§2.3) to promote the PRM to reward in both forward and backward sides. By focusing on step-level reasoning supervision tailored to clinical demands, our approach overcomes the limitations of prior models that rely on coarse labels or multiple-choice distillation unsuited for diverse clinical tasks. This results in a medical reasoning system optimized for evidence-based stepwise confidence accumulation, critical for trustworthy clinical decision support.

Extensive experiments on eleven clinical reasoning benchmarks demonstrate that MedS^3 achieves state-of-the-art performance (§4), outperforming both comparable-sized medical models and much larger general reasoning models, while maintaining superior interpretability and clinical task coverage. Table 1 highlights these advantages in robust long-chain reasoning and breadth of application.

In summary, our contributions are:

1. **First Step-Level Framework for Medical AI:** We introduce the first self-evolution framework that equips small-scale medical models with robust long-chain reasoning via step-level supervision, tailored for a wide range of clinical applications.

2. **Novel PRM Training Pipeline:** We propose a unique process reward model trained with soft dual-sided labels, which precisely evaluates each reasoning step by jointly predicting future rewards and assessing atomic step necessity, reflecting clinical reasoning's incremental confidence building.

3. **State-of-the-Art Clinical Reasoning Performance:** Our self-evolved system MedS^3 significantly surpasses all equal-parameter competitors and larger reasoning models across multiple clinical benchmarks, driven by fine-grained PRM-guided reasoning enhancement.

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 on the generated synthetic data with supervised learning and direct preference optimization (Rafailov et al., 2023).

3. **Process Reward Model (PRM)** V_θ (§2.3) which is fine-tuned with step-wise supervision using soft dual-side labels and assigns a

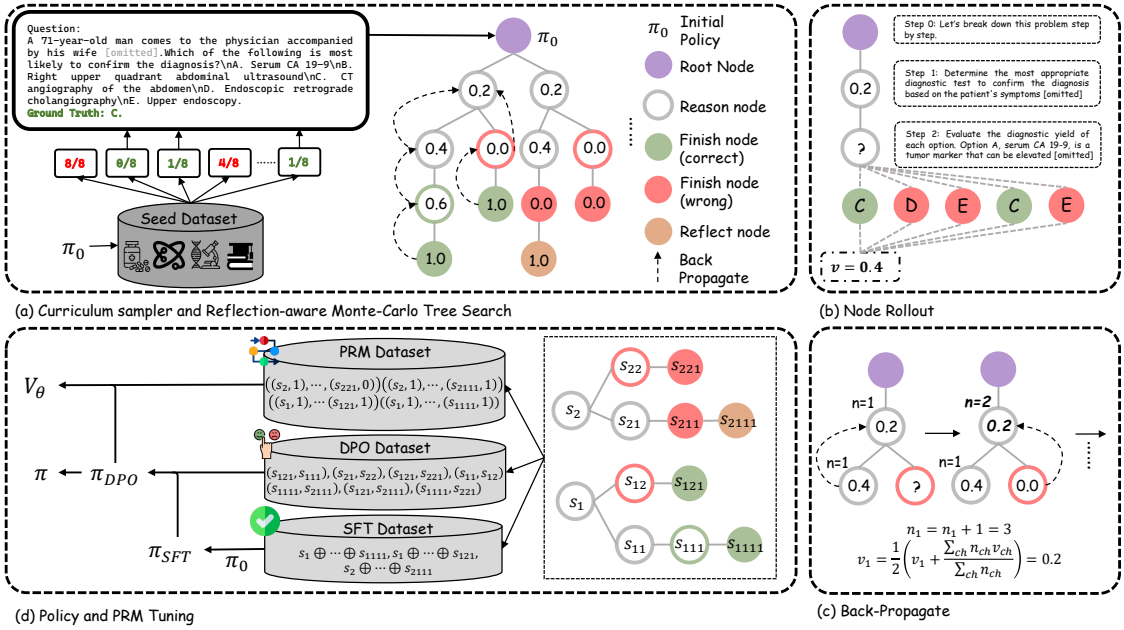


Figure 1: Overview of the construction of MedS^3 framework. (a) MedS^3 utilizes a Monte-Carlo Tree Search pipeline to self-generate step-by-step reasoning paths for each instance sampled in a curriculum manner. (b) During this process, MedS^3 uses result simulation to obtain the rollout value for each node; (c) After obtaining the child's rollout value, MedS^3 executes back-propagation to enable precise value prediction from deeper layers to transfer back to shallow nodes. (d) After gathering all correct and wrong finish nodes, we use SFT and DPO to optimize the policy model π and soft dual-side label to fine-tune a process reward model V_θ.

value in the range $[0, 1]$ to each reasoning step by a both forward and backward view.

4. **Iterative Training Pipeline** (§2.4) which consists of two MCTS evolution iterations and a curriculum data sampler.

The overall framework is presented in Fig. 1.

2.1 MCTS-guided Evolution

This algorithm builds upon an n-ary tree, where every root node is initialized as a reasoning start 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 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 preceding node of the current node T and γ is an exploration constant set as 2. For each intermediate node, we select its child node with the highest UCB value.

Node Expansion After reaching the candidate node T_c under the UCB criterion, we continue the reasoning trace 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 until deriving an answer for speeding up the exploration phase. On the other hand, for a wrong node, we allow one reflective action `Reflect` to elicit the introspection of the policy. 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 subsequent steps $\{s_{k+1,i} \mid i = 1, 2, \cdots, B\}$ 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, and x is the original input prompt.

[1]Prompts of `Finish`, `Reflect`, and `Reason` actions are illustrated in Appendix E

Node Rollout As the PRM is not yet accurate enough to serve as a reliable critic, node values are obtained using rollouts based on reasoning trajectories so far. Specifically, for a chosen unvisited node T_c at the k-th depth, we set a simulation budget $L = \min(L_{\min}, \frac{L_0}{k})$, 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 setting the budget, we prompt the policy model π_0 to directly output the answer L times under a specific prompt `AnsPrompt`:

$$a_c^l \sim \pi_0([s_0 \oplus s_1 \oplus \cdots \oplus s_k] \mid x_{\text{AnsPrompt}}), \quad (3)$$

where $l \in [1, L]$ and a_c^l is the l-th simulated answer. The average 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.

Backpropagation After the rollout stage, we conduct back-propagation starting from T_c till the root, updating all tree node values along the trace. Specifically, for an arbitrary node T_k, we propose to update its visits n_k and v_k as follows:

$$\begin{aligned} n_k &= n_k + 1 \\ v_k &= \frac{1}{2}\left(v_k + \frac{\sum_{ch} v_{ch} \cdot n_{ch}}{\sum_{ch} n_{ch}}\right), \end{aligned} \quad (4)$$

which considers both correctness and completeness for the evaluation of a reasoning step.

Termination of Search For balancing the exploration cost and optimization of policy and reward models, Therefore, we set two criteria to terminate the search process. First, once the total correct count in the tree exceeds a minimum correct count 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.

2.2 Policy Model Fine-tuning

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

$$\mathcal{L}_\pi = \frac{1}{L_k}\sum_{i=1}^{L_k} -\log p_\pi(y_i | x, y_{<i}), \quad (5)$$

where y_i is the i-th token of the reason trajectory and L_k is the total length of the trajectory. For the second iteration, we further add a step-level Direct Preference Optimization (DPO) to optimize the policy at the same reasoning budget:

$$\begin{aligned} \mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,P^+,P^-)\sim D_{\text{DPO}}} \log \sigma(r_\theta(x, P^+ \\ - r_\theta(x, P^-)), \end{aligned} \quad (6)$$

where $r_\theta(x, P) = \beta(\log \pi_\theta(P \mid x) - \log \pi_{ref}(P \mid x))$ is the reward and $D_{\text{DPO}} = \{(x, [s_0 \oplus s_1 \oplus \cdots \oplus s_k^+], [s_0 \oplus s_1 \oplus \cdots \oplus s_k^-]) \mid v_k^+ > v_k^-\}$. The DPO training is crucial for deriving a strong policy and PRM, which is elucidated in Table 3.

2.3 Soft Dual-side PRM Fine-tuning

Dataset Collection We first filter out those trees with only correct or incorrect leaves as these trajectories contain extreme value bias. For a `Finish` node T_k in a valid tree, its reasoning trace $[(s_1, v_1), (s_2, v_2), \cdots, (s_k, v_k)]$ is one training sample, where each reasoning step is concatenated by "`Step k:`" to form a complete reasoning trajectory. At the end of each reasoning step s_i (typically a `\n\n` token), the value v_i is used to derive the token label, which is learned by conditioning on all previous steps in an auto-regressive manner. As a result, the PRM training set is such $D_{V_\theta} = \{(x, [(s_1, v_1), (s_2, v_2), \cdots, (s_k, v_k)]) \mid x \in D_{seed} \wedge s_k \text{ is finish}\}$.

Learning objective Previous works in the math domain choose to directly learn the rollout value with Mean-Squared Error (Zhang et al., 2024a) or learn the pair-wise ranking preference (Guan et al., 2025). However, in our work, we propose to learn the prediction of the correctness probability of an intermediate step using a 2-class cross-entropy loss. The PRM V_θ is initialized from the tuned policy model for an aligned distribution, with the language model head replaced by a token classification layer with a cross-entropy loss for labeled tokens. 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, we deem that the rollout score as a soft label has a forward-only bias about reasoning correctness. Meanwhile, a rigorous and concise medical reasoning step, which cannot allow for exploring different solutions in a brute-force manner, is different from a math reasoning step (Yue et al., 2025). Therefore, a new step is valued highly only when it can both possibly derive a final answer and improve the correctness of the reasoning trajectory deterministically. As

a result, we design a dual-side label y_i for step i using its soft Q-value obtained during MCTS as

$$y_i = \begin{cases} \lceil v_i - \beta \cdot \max(0, v_{i-1} - v_{i+1}) \rceil & v_i < v_{i-1} \\ \lceil v_i \rceil & otherwise \end{cases} \tag{7}$$

This learning objective encourages PRM to simultaneously look ahead and back to judge the current step and penalize random trials except for reflective actions. Based on these, we optimize V_θ using the following loss function:

$$\mathcal{L}_{V_\theta} = \mathbb{E}_{T_k \sim D_{V_\theta}} \sum_{i=1}^{k} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i), \tag{8}$$

where $\hat{y}_i$ is the predicted probability of the given step i and β is a hyperparameter set to 1.0 by a simple grid search (details in Table 5). This dual-sided soft-label training, not only prevents the learning of fuzzy labels (rollout value around 0.5) but also learns to judge a misleading step.

2.4 Training Pipeline

We perform two iterations for the seed dataset. For each iteration, we use **curriculum sampler**, which first prompts the policy model to perform the rejected-sampling on the training set, filtering those training instances with all-correct responses to enhance data efficiency. After that, we sample instances with the lowest pass@1 values during the rejected-sampling process, ensuring that the extremely hard problems (0 pass@1 score) are no more than one-third of the total samples. After that, we perform MCTS evolution on the seed data and update the policy model. At the end of the second evolution, we further enhance the policy with DPO and train the PRM using the second iteration's data.

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 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

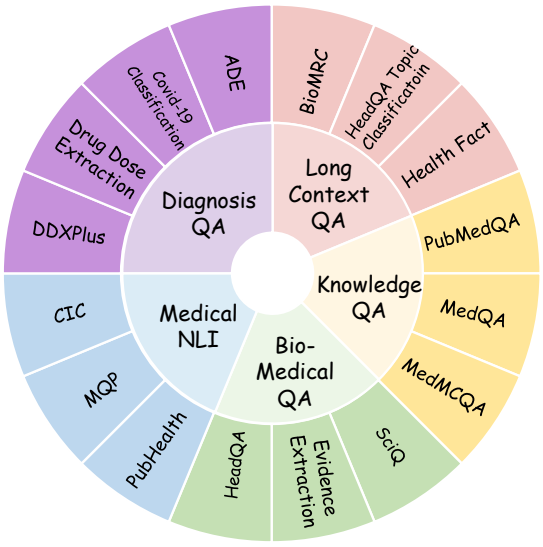


Figure 2: Overview of the used seed datasets.

medical datasets and divide them into five dimensions according to the task category. We show the five dimensions, i.e., clinical diagnosis QA, natural language inference, knowledge-intensive QA, long-context QA, biomedical QA and corresponding datasets in Fig. 2. The details about the definition of the five dimensions and the corresponding tasks can be found in Appendix F.

4 Experiments

In this section, we comprehensively evaluate 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 backbone of MedS^3. We select MedQA-5op (Jin et al., 2021), PubMedQA (Jin et al., 2019) without contexts, 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), DDX-Plus (Tchango et al., 2022) as the in-domain evaluation benchmarks, the medical subsets of MMLU (Hendrycks et al., 2021), BioASQ (Tsatsaronis et al., 2012) SEER Classification (Dubey et al., 2023) as the out-of-domain evaluation sets. The details of evaluation sets are presented in Appendix F and the hyperparameters of synthesis, self-training and evaluation are presented in Appendix G.

Models	MedQA†	MedMCQA†	PubMedQA†	BioASQ	MMLU	BioMRC	PubHealth	HealthFact	DDXPlus	DrugDose	SEER	Avg.
Large language models (>10B)												
GPT-4o-mini	75.81	67.58	47.80	83.01	83.79	66.85	59.14	65.24	54.00	73.91	54.54	66.52
GPT-3.5-turbo	59.31	58.12	37.40	74.11	71.11	56.22	57.84	67.85	39.05	86.96	73.61	61.96
QwQ-32B-preview	68.89	61.03	48.60	73.62	74.18	79.76	63.36	66.08	45.40	39.13	37.26	59.76
R1-Distill-Qwen32B	76.83	66.27	38.20	78.32	85.07	78.66	59.95	63.80	53.90	82.61	26.22	64.53
Small language models (<10B)												
Qwen2.5-7B	55.54	54.12	53.40	73.62	74.38	56.48	57.11	52.69	31.25	60.87	33.07	54.78
Llama3-8B	57.50	55.92	56.40	75.73	68.55	56.50	64.09	70.88	35.30	73.91	47.07	60.17
Llama3.1-8B	61.51	57.42	59.00	71.36	72.52	55.60	61.82	63.97	19.00	73.91	52.62	58.98
R1-Distill-Llama8B	50.12	48.89	46.60	70.55	68.42	53.49	55.73	62.04	36.10	69.57	31.71	53.93
Small Medical language models (<10B)												
MMedS-Ins	53.57	48.24	56.60	77.35	50.86	31.47	54.26	69.64	**97.53**	**95.65**	**97.93**	66.65
MedLlama3	55.85	59.36	66.40	84.63	70.08	47.97	62.39	68.10	22.50	69.57	50.69	59.78
Med42	50.20	49.70	55.40	74.76	61.43	57.26	59.14	**81.57**	31.35	65.22	37.14	56.65
OpenBioLLM	50.20	50.56	41.40	47.73	61.69	27.46	18.77	53.28	16.55	34.78	46.48	40.81
UltraMedical3-8B	68.89	61.82	51.60	80.58	75.08	45.18	66.13	72.73	36.70	60.87	24.55	58.56
UltraMedical3.1-8B	70.93	62.78	56.40	77.18	76.43	54.26	59.14	70.20	31.55	56.52	45.86	60.11
HuatuoGPT-o1	62.53	59.31	58.20	**87.70**	70.53	50.98	24.61	66.08	40.20	56.52	46.85	56.68
SFT on Seed	40.93	58.38	61.80	76.38	66.24	32.56	44.03	73.57	42.05	91.30	53.10	58.21
MedS^3 (ours)												
Iter 1	65.91	60.55	56.80	78.48	75.66	55.84	57.03	64.73	51.65	73.91	48.97	62.68
Iter 2	67.09	61.56	60.40	80.93	75.21	70.11	68.97	69.87	53.55	91.30	53.44	68.40
Iter 2 *w/ PRM*	**72.97**	**67.32**	**64.20**	81.39	**79.63**	**74.54**	**74.41**	76.18	62.40	91.30	59.80	**73.10**

Table 2: Experiment results in 11 medical datasets among four types of models. We highlight the best results with **bold** and underlines the second-best results among models with a similar size. "SFT on seed" denotes the variant of fine-tuning the policy on the seed data. † denotes the datasets on which most medical models have been trained.

Setting	MedQA	MedMCQA	PubMedQA	BioASQ	Med MMLU	BioMRC	PubHealth	HealthFact	DDX Plus	Drug Dose	SEER	Average
SFT Policy	64.69	61.46	57.80	80.26	75.98	63.28	63.44	64.23	52.65	78.26	48.85	64.63
w/ DPO	67.09	61.56	60.40	80.93	75.21	70.11	68.97	69.87	53.55	**91.30**	53.44	68.40
w/ H-S label	68.97	65.67	61.80	79.45	76.75	70.48	69.13	74.24	59.35	86.96	56.94	69.98
w/ H-D label	66.77	63.78	61.40	80.74	75.14	**78.13**	69.54	75.34	61.60	**91.30**	56.46	70.93
w/ S-D label	**72.97**	**67.32**	**64.20**	**81.39**	**79.63**	74.54	**74.41**	**76.18**	**62.40**	**91.30**	**59.80**	**73.10**
w/ SFT init. PRM	70.70	64.40	61.80	81.23	77.39	70.22	75.30	74.58	60.15	82.61	54.99	70.31

Table 3: Ablation study on each component of MedS^3 after the second iteration. "H-S" means hard single-sided label, "H-D" means hard dual-sided label, and "S-D" is soft dual-sided label used in MedS^3.

Baselines We choose the following two categories to serve as baselines: (1) LLMs, including GPT-3.5-turbo (OpenAI, 2022), GPT-4o-mini (OpenAI, 2023), QWQ-preview-32B (Qwen, 2024) and R1-Distill-Qwen32B (Guo et al., 2025); (2) Small Language models (<10B), including Llama 3 8B, Llama 3.1 8B (Dubey et al., 2024) and Qwen2.5 7B (Yang et al., 2024), R1-Distill-Llama8B (Guo et al., 2025) (3) 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), UltraMedical3-8B and UltraMedical3.1-8B (Zhang et al., 2024b) and HuatuoGPT-o1-8B (Chen et al., 2024). We also directly SFT the base model on the seed training set to illustrate no data contamination from the seed data. All the baselines are evaluated using CoT while MedS^3 *w/* PRM scores each response with the minimum step value and uses Best-of-N (N=32) to select the final response.

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

4.2 Main Results

We present the experiment results in Table 2, splitting into examination QA and clinical application tasks. 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 +14.12 average performance gains with respect to the base model in the overall assessment, which not only outperforms medical-oriented models but also general reasoning models. Specifically, the policy component has already achieved the state-of-the-art (SoTA) performance, based on which the soft dual-side PRM further brings an additional 4.7 points improvement.

Specifically, compared to HuatuoGPT-o1 and

Error Type	MedQA	MedMCQA	PubmedQA	Bioasq	Med MMLU	Biomrc	Pubhealth	Healthfact	DDX Plus	Drug Dose	SEER
Task	27.03	32.68	35.80	18.61	20.37	25.46	25.59	23.82	37.60	8.70	40.20
PRM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.29	0.00	4.35	0.00

Table 4: Comparison of error rates (lower is better) in task-level and PRM-level, where PRM works with no errors in most testbeds.

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 in-domain clinical tasks, such as SEER and DDX-Plus. 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 Ablation Study

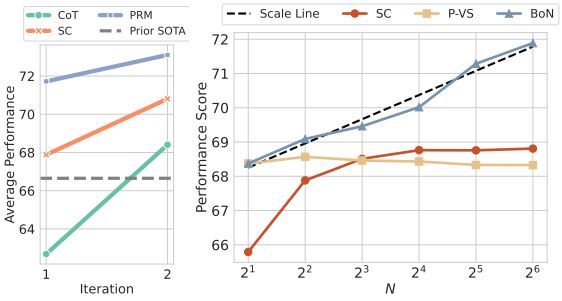
In this section, we validate the effectiveness of each sub-module of MedS^3. Starting from the SFT-tuned policy model, we compare the final performance with (1) *w/* DPO: use DPO to fine-tune the policy; (2) *w/* H-S label: conduct best-of-N evaluation using a PRM trained with hard single-sided label (Zhang et al., 2025); (3) *w/* H-D label: same as (2) but use hard dual-sided label (Wang et al., 2025) to train a PRM and (4) *w/* S-D label (ours): same as (2) but use soft dual-sided label proposed in MedS^3 to train a PRM. We also compare with (5) *w/* SFT init. PRM, which is same as (4) but initializes PRM with the SFT-tuned policy, to further show the significance of a PRM exposed to both positive and negative responses. Experiment results in Table 3 show that the DPO helps to greatly improve the policy model, especially in clinical tasks. Furthermore, innovatively determining the dual side label based on the MC estimation, our method is more robust and flexible than rule-based labels, and hence outperforms previous training objectives, confirming the necessity of holistic modeling of a PRM.

5.2 Reliability of PRM

Although our process reward model (PRM) is trained using rollout values—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 4, the majority of tasks exhibit a PRM error rate of zero, except for HealthFact and Drug Dose. 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.

5.3 Scaling of MedS^3

In this section, we present the improvements brought by the self-evolutionary framework in Fig. 3a, and those attributable to test-time scaling in Fig. 3b and Table 9. Specifically, we sample $n = 2, 4, 8, 16, 32, 64$ candidates for a prompt with a 1.0 temperature and compare the performance obtained through Best-of-N (BoN) (Lightman et al., 2023), PRM-guided Vote-Sum (P-VS; Wang et al. (2024)), as well as an SC baseline. We observe a great improvement in both the policy model and the PRM after a second evolution iteration, highlighting the efficacy of self-evolution. This suggests that the iterative MCTS process, where the model learns from its own refined outputs, leads to steadily increased improvements. Additionally, we find that test-time scaling further enhances MedS^3's reasoning performance as illustrated in Fig. 3b in an effective log-linear rate with little saturation. Together, these results highlight the benefits of both self-exploration during synthesis and self-supervision during inference, contributing to MedS^3's strong performance across diverse tasks.



(a) (b)

Figure 3: Scaling in (a) self-evolution iterations and (b) sampling numbers during test-time. Both the policy and PRM harvest consistent enhancement with self-evolution, and hence their cooperative system MedS^3 achieves a log-linear scaling rate with little saturation.

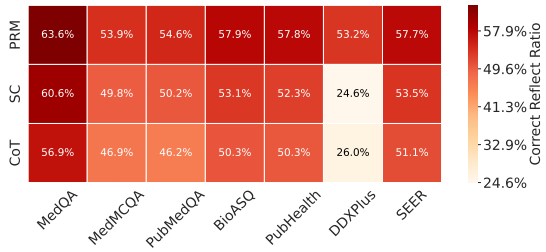


Figure 4: Reflective response ratio of MedS^3 across 7 representative datasets. Both the policy and PRM are reflection-aware to perform sequential test-time scaling.

5.4 Introspective Behavior

Reflection has been proved to be an effective scaling paradigm for enhancing LLM's test-time scaling capacity (Guo et al., 2025). Our MedS^3 introduced a `Reflect` node during synthesis and a soft dual-sided PRM to encourage correctly reflected responses, aiming to impart self-reflection behavior to the whole system. We manually define reflective tokens (`Wait, reevaluate, recheck, however, but`) and count the ratio of correct responses with these tokens on seven representative benchmarks in Fig. 4. We observe a steady increase in the occurring ratio from directly chain-of-thought prompting to leveraging PRM to conduct BoN evaluation, which indicates both the policy and PRM in MedS^3 has been imparted with self-reflection behavior. This further demonstrates that the PRM trained with the soft dual-sided label can correctly favor valuable responses with self-reflection.

5.5 Comparison of Slow-Thinking Styles

In this section, we compare three slow-thinking enhancement strategies, including 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. We use the first iteration dataset in §3 to implement RL, and use the officially released distillation dataset provided by Huang et al. (2025) to SFT the base model, and compare them with MedS^3 after the first evolution iteration. 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 a 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.

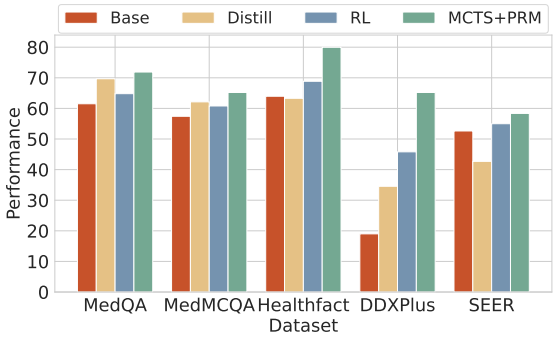


Figure 5: Three widely adopted methods to empower models with medical reasoning abilities. MCTS+PRM is the best among the three, making it the core of MedS^3.

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 covering 16 different realistic medical tasks, and use Monte-Carlo Tree Search to construct policy data and PRM data. We propose a new PRM learning objective – the soft dual-sided label, which enables the PRM to reward a step based on both future and past aspects, to produce credible long-chain reflective responses. Experiment results demonstrate that our model achieves superior performance on eleven downstream medical benchmarks, especially in realistic clinical ones, surpassing open-sourced models by a large margin with fewer parameters.

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 PRM learning objective: soft dual side label. However, it can be further improved via these strategies: (1) cooperate with reinforcement learning to empower the policy with the "aha-moment" (Guo et al., 2025) ability; (2) introduce more training samples to cover more medical reasoning scenarios; (3) conduct more evolution iterations to further improve the model.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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*.

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. Introducing OpenAI o3-mini. https://openai.com/index/openai-o3-mini/. Accessed: 2025-05-07.

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.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

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.

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.

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

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.

Teng Wang, Zhangyi Jiang, Zhenqi He, Wenhan Yang, Yanan Zheng, Zeyu Li, Zifan He, Shenyang Tong, and Hailei Gong. 2025. Towards hierarchical multistep reward models for enhanced reasoning in large language models. *arXiv preprint arXiv:2503.13551*.

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.

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*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? *Preprint*, arXiv:2502.12215.

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 sce-

narios 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 requires 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) improve 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 (Qwen, 2024).

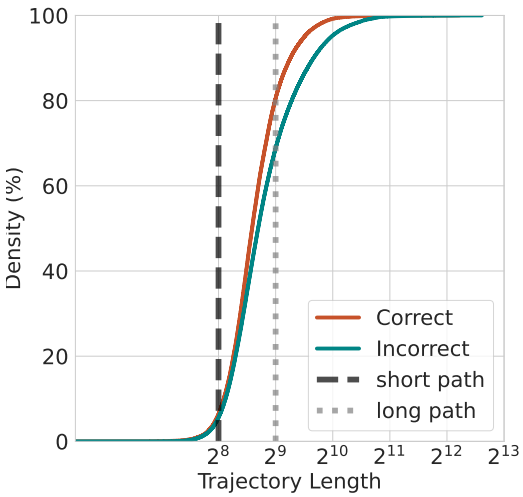


Figure 6: Trajectory length distribution of correct and incorrect sets of the evolved dataset.

Table 5: Grid search of β and corresponding loss in the dev set.

β	0.5	1	1.5	2
Dev loss	0.4293	0.4169	0.4194	0.538

B Statistics of the Evolved Dataset

In this section, we show the statistics of the evolved dataset after the second evolution, which is used to conduct our training of MedS^3. We plot the length distributions of trajectories in Fig. 6. Defining short sequences as below 256 and long sequences as above 512, we find that the evolved dataset contains about 20% long trajectories, which enables the policy model to generate reliable responses with more tokens. Moreover, we find that correct trajectories consume fewer tokens than incorrect ones, which aligns highly with Zeng et al. (2025).

C Further Experiments

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

C.1 Determination of β

We perform a simple grid search on a pre-defined dev set to find the most appropriate value o f β in Eq. 7. Specifically, we search β in the following list $[0.5, 1.0, 1.5, 2.0]$ and show the loss in the development set in Table 5. We determine β as 1 for its lowest loss. Although there might exist a more advanced configuration, we just set β to 1 as this is not our focus and we leave this for future work.

Model	MedQA	MedMCQA	PubHealth	HealthFact	BioMRC	PubMedQA	DDX Plus	Drug Dose	Average
GPT-4o-mini-ZS	75.81	67.58	59.14	65.24	66.85	47.80	54.00	73.91	63.79
GPT-4o-mini-FS	76.83	67.80	63.44	63.80	75.60	63.40	51.00	59.57	65.18
R1-Distill-Qwen-32B-ZS	76.83	66.27	59.95	63.80	78.66	38.20	53.90	82.61	65.03
R1-Distill-Qwen-32B-FS	76.36	81.40	59.06	64.14	87.40	41.80	51.40	60.87	65.30
MMedS-Ins-ZS	53.57	48.24	54.26	69.64	31.47	56.60	97.53	95.65	63.37
MMedS-Ins-FS	48.39	32.61	36.39	78.28	67.80	49.60	96.80	91.30	62.65
MedS3	72.97	67.32	74.41	76.18	74.54	64.20	62.40	91.30	**72.92**

Table 6: Comparison with prior Top-3 models prompted with the few-shot technique. Albeit certain improvements, these models still lag behind MedS^3 by a large margin. "ZS" means zero-shot while "FS" means few-shot.

Model	MedQA	MedMCQA	PubmedQA	Bioasq	Med MMLU	Biomrc	Pubhealth	Healthfact	DDX Plus	Drug Dose	SEER	Average
MMedS-Ins	54.36	52.76	63.60	72.82	59.26	23.95	59.95	74.66	**97.96**	**95.65**	**89.40**	67.67
Ultramedical3.1-8B	72.90	65.57	58.00	80.58	78.54	49.18	67.91	71.72	34.10	60.87	49.73	62.65
HuatuoGPT-o1	**75.96**	66.94	58.20	78.48	78.54	46.45	64.58	70.29	41.00	60.87	49.59	62.81
MedS^3	72.97	**67.32**	**64.20**	**81.39**	**79.63**	**74.54**	**74.41**	**76.18**	62.40	91.30	59.80	**73.10**

Table 7: Comparison with prior Top-3 models with similar model sizes prompted with self-consistency method. Albeit certain improvements, these models still lag behind MedS^3 by a large margin.

C.2 Comparison with Few-shot Prompted Models

Few-shot prompting is a widely adopted method to improve performance effortlessly. To this end, we compare MedS^3 with the 1-shot prompted Top-3 performing models, i.e., GPT-4o, R1-distill-32B and MMedS-Ins, in Table 2 where exemplars are provided by prompting o3-mini (OpenAI, 2024). We exclude benchmarks without training sets, including MMLU, SEER and BioASQ. Results shown in Table 6 indicate that in most cases, few-shot prompting does improve performance without any cost, whereas the medical-specialized model, MMedS-Ins, loses the in-context learning ability and underperforms its zero-shot variant. Meanwhile, few-shot prompted baselines still lag behind our self-evolved MedS^3 system, showing that the PRM-integrated medical reasoning framework is consistently powerful.

C.3 Comparison with SC Models

We also compare MedS^3 with baselines prompted with the Self-Consistency (SC) method, which is a simple yet efficient way to scale in a parallel manner. To maintain similar inference costs, we set the sampling number to 32 for models with similar size (<10B) and compare with the most powerful models before[3], namely HuatuoGPT-o1, UltraMedical3.1-8B and MMedS-Ins. Results in Table 7 illustrate that as a test-time scaling method, SC improves the already strong baselines by significant gains, while such improvements usually occur in traditional benchmarks like MedQA or MedMCQA. Their performances in clinical testbeds, like SEER or DDX-Plus, hardly show gains, which unveils some kind of overfitting problem. Therefore, optimized for both traditional examination and clinical usage, MedS^3 robustly achieves state-of-the-art performance overall.

[3]It is reasonable when MedS^3 outperforms the most leading baselines.

C.4 Backbone Selection

In this section, we investigate which backbone, a general LLM or a medical-specific LLM, is suitable for conducting self-evolution. We conduct one iteration of evolution using the same data of MedS^3 under UltraMedical3.1-8B and compare it with MedS^3 after the **first iteration** using the Best-of-N decoding method to save computational cost. The comparison shown in the upper half of Table 8 reveals that although UltraMedical3.1-8B improves compared to the initial policy, it still lags behind MedS^3 by a large margin. Delving into the generation, we find that UltraMedical3.1-8B suffers from endless generation, which stems from its lower instruction following ability compared to Llama 3.1 8B. On the other hand, medical backbones show no significant performance gains compared to the general model (UltraMedical3.1 8B 60.11 vs Llama 3.1 8B 58.98 in Table 2), while after optimized for certain benchmarks like MedQA, they have lower generalization ability than general models. Based on the above observations, we choose to use a general backbone with sufficient medical knowledge, i.e., Llama 3.1 8B, as the initial policy model.

Backbone	MedQA	MedMCQA	PubMedQA	BioASQ	Med MMLU	BioMRC	PubHealth	HealthFact	DDX Plus	Drug Dose	SEER	Average
UltraMedical3.1	**68.42**	58.20	58.00	**79.61**	73.16	49.40	68.07	71.38	49.20	86.96	51.40	64.89
Llama 3.1	67.64	**62.00**	**59.60**	79.13	**77.77**	**76.96**	**73.19**	**78.37**	**63.80**	**91.30**	**59.20**	**71.72**

Table 8: Comparison with UltraMedical-3.1-8B as the policy model. With superior instruction following ability and comparable medical knowledge, Llama-3.1-8B suits MedS^3 system to fulfill the self-evolution procedure.

Method	Iteration	MedQA	MedMCQA	PubmedQA	Bioasq	Med MMLU	Biomrc	Pubhealth	Healthfact	DDX Plus	Drug Dose	SEER	Average
BoN	2	68.97	64.04	62.00	79.45	76.43	73.68	72.14	73.57	58.00	86.96	55.45	70.06
	4	69.60	64.55	61.60	80.42	77.26	74.64	74.17	73.06	58.85	86.96	56.75	70.71
	8	70.54	64.57	62.60	81.07	77.83	74.75	74.09	72.64	59.90	86.96	56.61	71.05
	16	70.23	66.32	64.00	81.23	78.41	74.80	73.68	72.05	61.00	86.96	58.44	71.56
	32	72.97	67.32	64.20	81.39	79.63	74.54	74.41	76.18	**62.40**	91.30	59.80	73.10
	64	**73.37**	**67.65**	**66.00**	**81.72**	**79.37**	74.54	**74.90**	**78.28**	62.25	**91.30**	**60.79**	**73.65**
SC	2	65.67	61.49	60.60	77.02	73.73	71.09	68.48	70.79	56.45	91.30	52.59	68.11
	4	67.09	63.11	60.40	78.80	75.72	73.23	70.59	76.18	57.35	91.30	56.32	70.01
	8	67.40	63.71	60.60	80.42	76.30	73.82	70.11	77.61	57.65	91.30	57.48	70.58
	16	68.42	63.73	60.80	80.42	76.43	73.70	70.11	77.69	58.05	91.30	58.27	70.81
	32	67.64	63.52	60.60	80.26	76.55	73.98	70.59	**78.28**	57.90	91.30	58.25	70.81
	64	67.79	63.45	60.80	80.26	76.75	73.98	70.76	77.86	58.10	91.30	58.33	70.85
P-VS	2	68.97	64.04	62.00	79.45	76.43	73.68	72.14	73.57	58.00	86.96	55.45	70.06
	4	68.97	63.85	60.40	80.26	75.91	74.38	71.57	75.25	57.70	86.96	57.37	70.24
	8	68.34	63.95	61.00	80.74	76.55	74.54	70.11	76.60	55.40	86.96	57.37	70.14
	16	68.81	63.88	60.80	81.39	77.07	74.99	70.27	75.67	53.10	86.96	58.33	70.12
	32	68.66	63.81	61.20	80.74	76.81	74.88	71.16	74.41	53.65	82.61	57.99	69.63
	64	68.19	63.71	61.00	80.58	77.39	**74.88**	71.24	74.41	53.70	82.61	58.16	69.62

Table 9: Full table of test-time scaling using PRM with different evaluation methods.

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

2. 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.

E Prompt Template

We show the prompt used to synthesize reasoning data in Fig. 7, Fig. 8, and Fig. 9.

F Dataset Details

In this section, we elucidate the seed dataset and the evaluation sets. We also clearly denote the involved dataset's usage during training and evaluation and their corresponding category in Table 10. 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 HealthFact (Kotonya and Toni, 2020)

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), and PubMedQA (Jin et al., 2019).

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 contain PubHealth (Kotonya and Toni, 2020), Medical

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 7: 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 8: Finish template

Reflect 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: [omitted]

Step k: [omitted]. The answer is C.

Step k+1: Wait, the previous answer maybe incorrect and I need to reconsider
other options.
```

Figure 9: Reflect template

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, Adverse Drug Event Classification (Huynh et al., 2016; Wu et al., 2024b) and DDX-Plus (Tchango et al., 2022)..

The descriptions of each training and evaluation dataset 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, 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 express 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 two given 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.

15. SEER (Dubey et al., 2023) is purposed for treatment planning because it contains key clinical variables that directly inform therapy decisions (e.g., tumor size, nodal status, hormone receptor status). LLMs must choose the most appropriate suggestion from the following list ['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'] based on patient summarization, simulating real-world tumor board decisions.

G Hyperparameters

G.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.

G.2 Self-Training of Policy and PRM

We use 8xNVIDIA A100 GPUs and the overall training consumes 14h.

Policy tuning We use trl[4] as the training framework. We first 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

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

Category	Dataset	Train	Test
Diagnosis QA	ADE	Yes	No
	Covid-19 CLS	Yes	No
	DrugDose	Yes	Yes
	DDXPlus	Yes	Yes
	SEER	No	Yes
Medical NLI	PubHealth	Yes	Yes
	CIC	Yes	No
	MQP	Yes	No
Long Context QA	BioMRC	Yes	Yes
	HealthFact	Yes	Yes
	HeadQA Topic CLS	Yes	No
BioMedical QA	HeadQA	Yes	No
	Evidence Extraction	Yes	No
	SciQ	Yes	No
Knowledge QA	MedQA	Yes	Yes
	MedMCQA	Yes	Yes
	PubMedQA	Yes	Yes
	MMLU	No	Yes
	BioASQ	No	Yes

Table 10: Medical datasets usage during training and evaluation. "CLS" denotes classification.

set to 128 and the learning rate is set to 1e-6. After that, we use `DPOTrainer` to further fine-tune the policy model. We set the learning rate to 5e-8 and the batch size to 128.

PRM tuning We use `PRMTrainer` of trl to train the PRM model. We use LoRA to fine-tune the PRM, where the lora rank is set to 32 and lora alpha set to 64. The learning rate is set to 5e-5. For a single step s_k, the input for PRM is the concatenation of all steps up to the current step, namely:

$$P = s_0 \oplus s_1 \oplus \cdots \oplus s_k \quad (9)$$

$$\hat{y} = V_\theta(P; x) \quad (10)$$

This input models a step's value with causal relationships between steps, preventing local optima learning.

G.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 $N = 32$ samples for a given problem, and we select the one whose answer appears most times among the N outputs. 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-processing 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.

G.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[5] 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 11.

Table 11: 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

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

H Best-of-N Details

In this section, we elucidate the fast inference using Best-of-N (BoN) evaluation with the PRM. Specifically, the policy model generates N responses $\{y_i \mid i \in [1, N]\}$ simultaneously using the inference engine (vLLM; Kwon et al. (2023)). After that we split each response y_i into K steps: $[s_i^1, s_i^2, \cdots, s_i^k]$ where $k \in [1, K]$ using predefined step separator `Step`. We form the step input to the PRM (x, S_i^k) as $S_i^k = s_i^1 \oplus s_i^2 \oplus \cdots \oplus s_i^k$ and obtain each step's value $v_i^k = \mathrm{softmax}(V_\theta(x; S_i^k))$ using a forward pass. The value is the probability of predicting the step to be correct, which is in the range $[0, 1]$. Finally, the whole response value $v_i = \min_{k=1}^{K} v_i^k$ is taken from the minimum value of each step and we select the response with the highest value.

I AI Assistant Usage

We mainly use AI assistants to polish our expressions.

J Case Study

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 initial clinical scenario and physical examination findings. The patient is an asymptomatic 47-year-old executive undergoing a routine check-up. The physical examination reveals a mid-systolic murmur heard at the 2nd left intercostal space radiating to the carotids. This murmur is suggestive of a pulmonary valve issue, but given the radiation to the carotids, it's more likely to be a semilunar valve issue, and the closest one would be the aortic valve, however the classic exam for aortic stenosis is the chest pressure, and the one described would be more indicative of a pulmonary valve issue but in the given location it's likely aortic stenosis. However, the key piece of information here is the physician's instruction to have the patient stand from a supine position while continuing to auscultate the heart.\n\nStep 2: Understand the effect of changing body position on cardiovascular physiology. When a person transitions from a supine (lying down) position to a standing position, several physiological changes occur:\n\n- Venous blood pools in the lower extremities due to gravity, reducing venous return to the heart.\n- Cardiac output decreases due to reduced preload.\n- Heart rate may increase to maintain cardiac output (baroreceptor reflex).\n- Systemic vascular resistance may increase to compensate for the decreased cardiac output.\n\nThe specific maneuver instructed by the physician, where the patient stands from a supine position while continuing to auscultate the heart, will lead to a decrease in venous return and, consequently, a decrease in left ventricular preload. This change in preload can affect the intensity of heart sounds, particularly those related to ventricular filling.\n\nStep 3: Relate the physiological changes to the potential effects on the heart sounds and the murmur in question. In the case of a mid-systolic murmur, as described, the intensity of the murmur is more likely related to the flow across the valve during systole. However, in this scenario, the key is to understand how the change in body position affects the physiology of the heart and the likelihood of certain effects on heart sounds and pressures.\n\nGiven the change in position, the most likely effect on the pressures and heart sounds would be a decrease in the diastolic filling time due to decreased venous return and thus a decrease in left ventricular end-diastolic pressure and volume, however the best answer given would be a decrease in diastolic filling time due to decreased preload, although the best answer regarding a direct relationship between the two isn't given in the question options.\n\nHowever, we must find the closest relation. If we consider the decrease in preload, and looking at the options:\n\n- A. An increase in right atrial pressure is not directly related to the change in body position as described.\n- B. An increase in left ventricular end-diastolic pressure is the opposite of what would be expected with decreased preload.\n- C. An increase in pulmonary capillary wedge pressure is also not directly related to the change in body position.\n- D. A reduction in the slope of the pacemaker potential is related to the effect of decreased preload on the sinoatrial node but is more related to the baroreceptor reflex rather than the direct effect of decreased preload on the ventricles.\n- E. A reduction in diastolic filling time is the closest option, as it is directly related to the decrease in preload due to the change in body position.\n\n**The answer is E.**

Figure 10: 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: 45 \n\nSex: M \n\nInitial evidence: Do you have a cough that produces colored or more abundant sputum than usual? Yes \n\nEvidence: Do you have pain somewhere, related to your reason for consulting? Yes;Characterize your pain: Sensitive;Do you feel pain somewhere? Lower chest;Do you feel pain somewhere? Side of the chest(r);Do you feel pain somewhere? Breast(r);Do you feel pain somewhere? Posterior chest wall(r);Do you feel pain somewhere? Posterior chest wall(l);How intense is the pain? 2;Does the pain radiate to another location? Nowhere;How precisely is the pain located? 5;How fast did the pain appear? 3;Do you have a cough that produces colored or more abundant sputum than usual? Yes;Do you smoke cigarettes? Yes;Do you have a chronic obstructive pulmonary disease (COPD)? Yes;Do you have a cough? Yes;Have you traveled out of the country in the last 4 weeks? N;Have you noticed a wheezing sound when you exhale? Yes.

Llama3.1-8B:
acute copd exacerbation / infection. \nThe answer is based on the patient's history of COPD, smoking, and symptoms of a productive cough, colored or more abundant sputum, and wheezing sound when exhaling..

MedS[3]
Step 0: Let's break down this problem step by step.\n\nStep 1: Identify key symptoms and risk factors.\n- The patient has a cough producing colored or more abundant sputum than usual.\n- He experiences pain in the lower chest, right side of the chest, right breast, and posterior chest wall (both sides).\n- The pain is sensitive, with an intensity of 2/10 and appeared within 3 days.\n- He has a history of smoking and COPD.\n- He also reports a wheezing sound when exhaling.\n\nStep 2: Consider possible diagnoses based on symptoms and risk factors.\n- Given the patient's history of smoking and COPD, chronic bronchitis or acute exacerbation of COPD could be a possibility.\n- However, the presence of colored sputum and pain in the chest and breast areas could suggest a more complex condition.\n- The patient's symptoms, such as cough, sputum production, and wheezing, are consistent with a respiratory condition.\n\nStep 3: Evaluate the likelihood of each possible diagnosis.\n- Considering the patient's history of smoking and COPD, bronchitis (chronic or acute) is a strong possibility.\n- However, the presence of colored sputum and pain in the chest and breast areas could also suggest bronchiectasis, a condition characterized by damaged airways that can produce excess mucus.\n- Other conditions, such as pneumonia or pulmonary embolism, could also present with similar symptoms, but they are less likely given the patient's history and the absence of other risk factors.\n\nStep 4: Based on the patient's symptoms and risk factors, the most likely diagnosis is bronchitis, likely an acute exacerbation of his chronic COPD. However, considering the presence of colored sputum and pain, bronchiectasis is also a possibility, but given the options provided, **the best answer is bronchitis.**

MedS[3]-PRM:
Step 0: Let's break down this problem step by step.\n\nStep 1: Identify the key symptoms and risk factors. The patient presents with a cough producing colored or more abundant sputum than usual, pain in the lower chest, right side of the chest, right breast, and posterior chest wall (both sides), which is sensitive and appeared within 3 days. He also reports a history of smoking and COPD.\n\nStep 2: Consider the potential causes of these symptoms. Given the history of smoking and COPD, chronic bronchitis or acute exacerbation of COPD could be a possibility. However, the presence of colored sputum and localized chest pain may suggest a more specific condition. The patient's symptoms could also be consistent with pneumonia or bronchiectasis, but these conditions typically present with more systemic symptoms and findings on physical examination or imaging.\n\nStep 3: Assess the patient's risk factors for specific conditions. The patient's smoking history and COPD increase his risk for chronic obstructive pulmonary diseases (COPDs) and pulmonary infections. However, the patient's age and sex also increase his risk for other conditions such as lung cancer.\n\nStep 4: Based on the patient's symptoms and risk factors, the most likely diagnosis is an acute exacerbation of COPD or chronic bronchitis. However, the presence of colored sputum and localized chest pain also raises the possibility of bronchiectasis. Considering the patient's history of smoking and COPD, the most likely diagnosis among the options provided is bronchiectasis or acute bronchitis/bronchiolitis but I see that the best answer here is an exacerbation of the COPD which is not listed but we do have 'bronchitis' which could be correct in this context.\n\n**The answer is bronchitis.**

Figure 11: 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 determine the stage and characteristics of the breast cancer. The patient is a 73-year-old female with a tumor size of 16mm, which is likely to be stage I or stage II breast cancer. The patient has estrogen receptor-positive (ER+) breast cancer, but HER2-negative and progesterone receptor-negative (PR-) status. There is no information on lymph node involvement, but 8 nodes were examined.\n\nStep 2: Assess the patient's overall health and treatment goals. The patient is 73 years old, and the cause of death is listed as diseases of the heart. This suggests that the patient may have had underlying cardiovascular conditions that could impact her ability to tolerate certain treatments.\n\nStep 3: Consider the standard treatment options for early-stage breast cancer. For a patient with ER+ breast cancer, hormone therapy is often a recommended treatment option. However, the patient's age and cardiovascular conditions may impact the choice of treatment.\n\nStep 4: Evaluate the radiation therapy options. Given the patient's age and cardiovascular conditions, it may be more appropriate to avoid invasive surgery and radiation therapy. However, the patient's tumor size and lymph node status are not sufficient to rule out the need for radiation therapy entirely.\n\nStep 5: Determine the most appropriate next step in treatment based on the patient's characteristics and overall health. Considering the patient's age, cardiovascular conditions, and ER+ status, a more conservative approach may be preferred. However, there is no clear indication that the patient received adequate radiation therapy. A more conservative option would be to provide hormone therapy and monitor the patient's condition. But among the provided options, \"No radiation and/or cancer-directed surgery\" seems like the best answer given the lack of information on whether surgery was performed and the patient's overall health. \n**The answer is 'No radiation and/or cancer-directed surgery'.**

Figure 12: Cases of SEER