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

Shuyang Jiang[1, 3], **Yusheng Liao**[2, 3], **Zhe Chen**[2, 3], **Ya Zhang**[2, 3, *], **Yanfeng Wang**[2, 3], **Yu Wang**[2, 3, *]
[1]Fudan University
[2]School of Artificial Intelligence, Shanghai Jiao Tong University
[3]Shanghai Artificial Intelligence Laboratory
shuyangjiang23@m.fudan.edu.cn, {liao20160907,chenzhe2018,ya_zhang, wangyanfeng622, yuwangsjtu}@sjtu.edu.cn

Abstract

Medical language models face critical barriers to real-world clinical reasoning applications. However, mainstream efforts, which fall short in task coverage, lack fine-grained supervision for intermediate reasoning steps, and rely on proprietary systems, are still far from a versatile, credible and efficient language model for clinical reasoning usage. To this end, we propose MedS^3, a self-evolving framework that imparts robust reasoning capabilities to small, deployable models. Starting with 8,000 curated instances sampled via a curriculum strategy across five medical domains and 16 datasets, we use a small base policy model to conduct Monte Carlo Tree Search (MCTS) for constructing rule-verifiable reasoning trajectories. Self-explored reasoning trajectories ranked by node values are used to bootstrap the policy model via reinforcement fine-tuning and preference learning. Moreover, we introduce a soft dual process reward model that incorporates value dynamics: steps that degrade node value are penalized, enabling fine-grained identification of reasoning errors even when the final answer is correct. Experiments on eleven benchmarks show that MedS^3 outperforms the previous state-of-the-art medical model by +6.45 accuracy points and surpasses 32B-scale general-purpose reasoning models by +8.57 points. Additional empirical analysis further demonstrates that MedS^3 achieves robust and faithful reasoning behavior.

Code — https://github.com/pixas/MedSSS
Extended version — https://arxiv.org/pdf/2501.12051

1 Introduction

Large Language Models (LLMs) have demonstrated significant potential in the medical domain (Singhal et al. 2023; Nori et al. 2023), supporting tasks from clinical note generation (Biswas and Talukdar 2024; Jung et al. 2024) to precise diagnosis (Tu et al. 2025; Liao et al. 2024). Despite these advances, accurate reasoning is steadily fundamental to clinical decision-making, where diagnostic and treatment recommendations must be grounded in coherent, evidence-based logic chains (Cabral et al. 2024; Tordjman et al. 2025). Increasing efforts to enhance reasoning capabilities through chain-of-thought (Wei et al. 2022), preference learning (Rafailov et al. 2023) and reinforcement learning (Guo et al. 2025) highlight that correct answers alone are insufficient without trustworthy reasoning processes.

While existing approaches have demonstrated notable performance, two challenges persist. First, the training data used in many studies mostly consist of multiple-choice problems (Huang et al. 2025b,a), which lacks sufficient diversity and scale, and hence limits model robustness across different domains. Second, a growing dependence on large-scale proprietary models (Huang et al. 2025b) introduces practical and ethical considerations. Although models distilled from these hyper-scale teachers achieve strong performance, they inherit the unverifiability and potential hallucinations (Xu, Jain, and Kankanhalli 2024) of their teachers, offering little control over reasoning faithfulness. Moreover, the reliance on distillation from external resources would lead to uncontrollable privacy protection for real-world applications. These challenges highlight a core problem: how to efficiently induce robust, interpretable, and stepwise supervision reasoning in small-scale medical models without relying on proprietary models or noisy synthetic supervision.

To bridge this gap, we propose MedS^3, a self-evolving medical reasoning framework that enables small models to iteratively improve through Monte Carlo Tree Search (MCTS)-guided exploration and rule-verified refinement. Starting from a diverse, curriculum-sampled dataset spanning five medical domains and 16 medical datasets, MedS^3 generates reasoning trajectories with explicit node value estimates, allowing selection of high-quality paths for policy model bootstrapping via reinforcement fine-tuning and preference learning. Crucially, we introduce a soft dual-sided process reward model that labels intermediate steps not only by potential correctness but also by value consistency—penalizing steps that degrade node value and marking them as incorrect if the adjusted value falls below zero. This enables faithful supervision even in trajectories with correct final answers. Table 1 highlights these advantages in robust long-chain reasoning and breadth of application.

Extensive experiments on eleven clinical reasoning benchmarks and three out-of-domain datasets demonstrate that MedS^3 achieves state-of-the-art performance, outperforming both comparable-sized medical models and much larger general reasoning models, while maintaining superior interpretability and clinical task coverage. In summary, our

*Corresponding Authors

Models	Without Close-sourced Teacher	Diverse Clinical Coverage	Small Size	Reasoning Specialized	Process Supervision
UltraMedical	✗	✓	✓	✗	✗
HuatuoGPT-o1	✗	✗	✓	✓	✗
O1-journey Part 3	✗	✗	✗	✓	✗
m1-7B-32K	✗	✗	✓	✓	✗
MedS^3	✓	✓	✓	✓	✓

Table 1: Comparison of MedS^3 with other medical models. MedS^3 supports flexible inference-time scaling on resource-constrained devices, as well as process reward-guided decoding algorithms without supervision from large proprietary models.

contributions are:

1. **Pioneering Step-Level Framework for Medical AI:** We introduce a 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 and fewer hallucinations.
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 presented in Fig. 1. It is structured into four 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 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.

2.1 MCTS-guided Evolution

This algorithm builds upon an n-ary tree, where every root node is initialized as a multi-step reasoning start s_0 ="Let s break down this problem step by step.". 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, Björnsson, and Saito 2008) as the criterion to select a child T, which is as follows:

$$\mathrm{UCB}_T = v_C + \gamma\sqrt{\frac{\ln n_{T_{parent}}}{n_T}}, \tag{1}$$

where T_{parent} is the preceding node of the current node T, n_T is the node visiting count, v_C is the node's value obtained by node rollout and updated by back-propagation, and γ is an exploration constant set as 2. For each parent, 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 exploration. 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 layer of 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` node:

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

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.

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 $L = \max(3, \frac{L_0}{k})$ where $L_0 = 15$, 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`:

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

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.

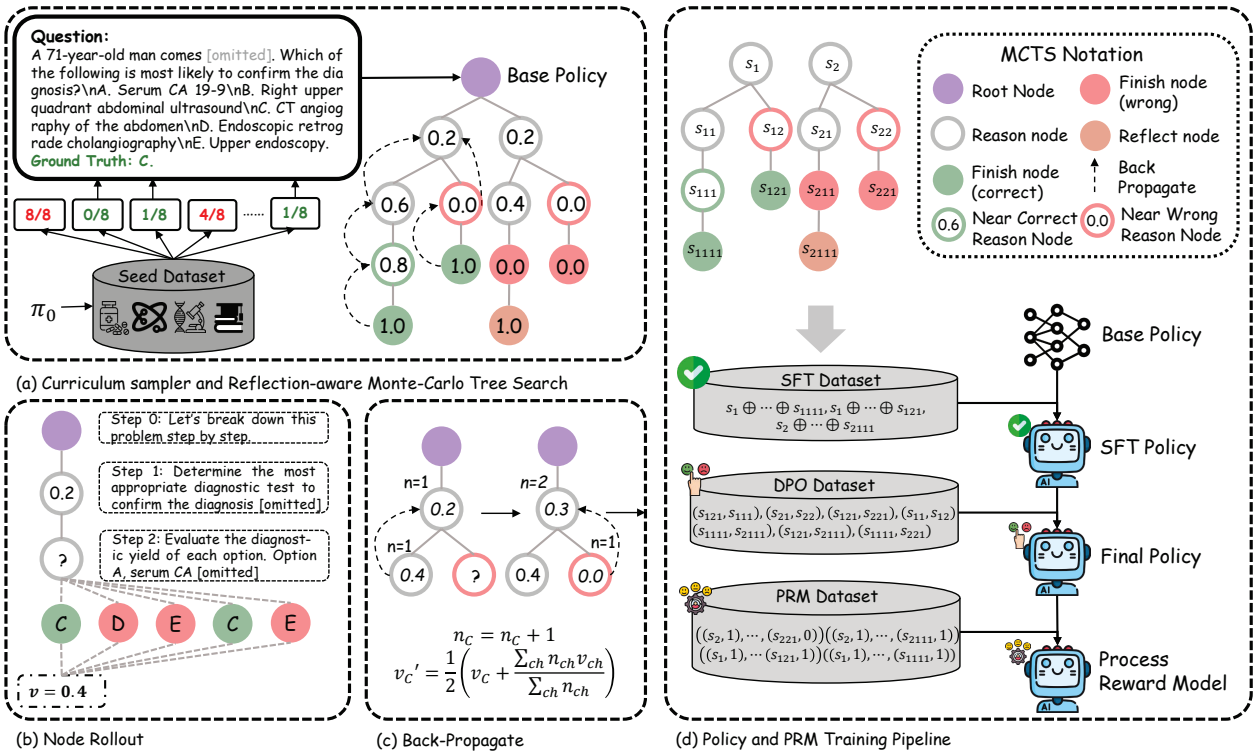


Figure 1: Overview of the construction of MedS^3 framework. (a) MedS^3 utilizes a Monte-Carlo Tree Search pipeline to self-generate step-by-step reasoning paths for each instance sampled in a curriculum manner. (b) During this process, MedS^3 uses result simulation to obtain the rollout value for each 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 the exploration finishes, we use SFT and DPO to optimize the policy model and soft dual-side label to fine-tune the process reward model.

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:

$$n_k = n_k + 1$$
$$v_k = \frac{1}{2}\left(v_k + \frac{\sum_{ch} v_{ch} \cdot n_{ch}}{\sum_{ch} n_{ch}}\right), \tag{4}$$

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

Termination of Search To balance the exploration cost and optimization of policy and reward models, we set two criteria to terminate the exploration. First, once the total correct count in the tree exceeds a minimum correct count $\tau = 3$, 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 first leverages the correct leaves s_l whose values equal 1 and corresponding reasoning trajectories gathered before: $D_\pi = \{(x, [s_0 \oplus s_1 \oplus \cdots \oplus s_l]) \mid v_l = 1\}$. These correct reasoning traces are fine-tuned to deduce a self-improved policy model:

$$\mathcal{L}_\pi = -\mathbb{E}_{(x,y)\sim D_\pi} \log p_\theta(y \mid x), \tag{5}$$

where $y = [s_0 \oplus s_1 \oplus \cdots \oplus s_l]$ is the whole trajectory. For the second iteration, we further add a step-level Direct Preference Optimization (DPO) to optimize the policy:

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

where $r_\theta(x, P) = \beta(\log \pi_\theta(P \mid x) - \log \pi_{ref}(P \mid x))$ is the reward and $D_{\text{DPO}} = \{(x, [s_0 \oplus s_1 \oplus \cdots \oplus s_k^+], [s_0 \oplus s_1 \oplus \cdots \oplus s_k^-]) \mid v_k^+ > v_k^-\}$. DPO training is crucial for deriving a strong policy and PRM, which is elucidated in Table 4.

2.3 Soft Dual-side PRM Fine-tuning

Dataset Collection We first filter out trees with only correct or incorrect leaves to avoid extreme value bias. For a valid `Finish` leaf T_l, its reasoning trace $[(s_1, v_1), \cdots, (s_l, v_l)]$ 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 $D_{V_\theta} = \{(x, [(s_1, v_1), (s_2, v_2), \cdots, (s_l, v_l)]) \mid x \in D_{seed} \wedge s_l \text{ is finish}\}$.

Learning objective Instead of fitting the node value (Zhang et al. 2024a) or learning the pair-wise ranking preference (Guan et al. 2025), we choose to use a binary cross-entropy loss to optimize the PRM for its stability. Although Zhang et al. (2025) suggests that the PRM label should be set to True 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. A wrong intermediate step is still possible to derive a correct answer given correct prefixes, but such hallucinations are not what medical reasoning desires. 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

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

This learning objective encourages PRM to simultaneously look ahead and back to judge the current step and penalize intermediate errors except for valid reflection. Based on these, we optimize V_θ using the following loss function:

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

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 Appendix C.1). 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 average accuracy values during the rejected-sampling process, ensuring that the extremely hard problems (0 accuracy 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 queries (Huynh et al. 2016). However, previous works mainly 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 broader range of clinical tasks, we curate a training corpus from 16 existing public medical datasets and divide them into five dimensions according to the task category. The five dimensions, i.e., clinical diagnosis QA, natural language inference, knowledge-intensive QA, long-context QA, biomedical QA and corresponding datasets are shown in Fig. 2.

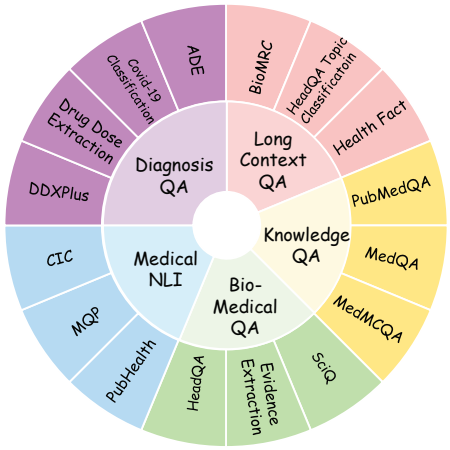


Figure 2: Overview of the used seed datasets.

4 Experiments

In this section, we comprehensively evaluate MedS^3 on 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, Umapathi, and Sankarasubbu 2022), PubHealth (Kotonya and Toni 2020), BioMRC (Pappas et al. 2020), HealFact Classification (Kotonya and Toni 2020), Drug Dose Extraction (Huynh et al. 2016), DDXPlus (DDX+; Tchango et al. (2022)), the medical subsets of MMLU (Hendrycks et al. 2021), BioASQ (Tsatsaronis et al. 2012) SEER Classification (Dubey et al. 2023) as the evaluation sets.

Baselines We choose the following three categories to serve as baselines: (1) LLMs, including GPT-3.5-turbo (OpenAI 2022), GPT-4o-mini (Achiam et al. 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 (Yonsei 2024), Med42 (Christophe et al. 2024), OpenBioLLM (Ankit Pal 2024), UltraMedical3-8B and UltraMedical3.1-8B (Zhang et al. 2024b), m1-7B-32K (Huang et al. 2025a) and HuatuoGPT-o1-8B (Chen et al. 2025). 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.

Models	MedQA	MedMCQA	PQA.	BioASQ	MMLU	BioMRC	PubH.	HFact.	DDX+†	DrugD.†	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)												
MedLlama3	55.85	59.36	66.40	<u>84.63</u>	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
m1-7B-32K	70.70	61.85	48.60	77.83	78.35	52.93	56.70	61.62	29.15	69.57	56.70	60.36
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
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**	<u>76.18</u>	**62.40**	**91.30**	**59.80**	**73.10**

Table 2: Experiment results in 11 in-domain datasets. We highlight the best results with **bold** and <u>underlines</u> the second-best results among models with a similar size. 'PQA.' denotes 'PubMedQA', 'PubH.' denotes 'PubHealth', 'HFact.' denotes 'HealthFact', and 'DrugD.' denotes 'DrugDose'. † denotes that the ground truth is not a simple choice index.

Model	MedCalc	MedXpert	RDC
GPT-4o-mini	29.80	15.43	37.80
HuatuoGPT-o1	21.97	16.04	16.20
UltraMedical-3.1-8B	15.19	16.12	21.80
R1-Distilled-Llama-8B	11.94	12.65	13.60
MedS^3	23.69	16.20	33.20
MedS^3 *w/* PRM	**30.66**	**16.44**	**41.20**

Table 3: Out-of-domain comparison between MedS^3 and previous state-of-the-art models. MedS^3 achieves great generalization ability on both policy and process reward models.

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 multiple-choice problems; while such superiority falls short on out-of-distribution real-world clinical benchmarks (DDXPlus or SEER), resulting in a sub-optimal overall performance compared to the general LLM–Llama3-8B. In contrast, our MedS^3 is tailored for universal medical applications and hence achieves the best overall 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, outperforming both medical-oriented models and general reasoning models. After two iterations, the policy model individually achieved the state-of-the-art (SoTA) performance, based on which the soft dual-side PRM further brings an additional 4.7 points improvement. Notably, unlike previous methods that rely on large volumes of multiple-choice queries and consequently suffer from over-fitting, MedS^3 achieves robust reasoning improvements, demonstrating that as few as 1,000 high-quality seed examples per task are sufficient for clinical reasoning.

4.3 Generalization to Out-of-domain Tasks

To validate the efficacy of MedS^3 on real-world tasks with little labeled data, we select the most frontier models, including GPT-4o-mini, R1-Distill-Llama8B, HuatuoGPT-o1 and UltraMedical3.1-8B as the competitors and further compare MedS^3 on MedCalc (Khandekar et al. 2024), MedXpert (Zuo et al. 2025) and the rare disease confirmation (RDC) part sourced from PMCPatients (Zhao et al. 2023). Experiment results in Table 3 illustrate that both the policy and the PRM are applicable to unseen problems and the reasoning manner incentivized by self-evolution is sufficient for both clinical rare disease reasoning and more challenging reasoning scenarios.

5 Analysis

5.1 Ablation Study

In this section, we validate the effectiveness of each submodule 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

Setting	MedQA	MedMCQA	PQA.	BioASQ	MMLU	BioMRC	PubH.	HFact.	DDX+	DrugD.	SEER	Avg.
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 4: 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.

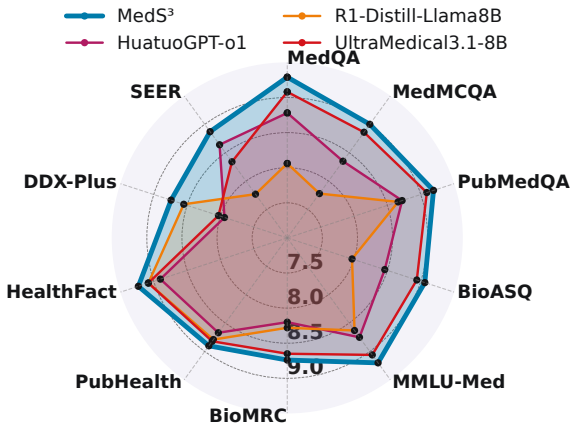
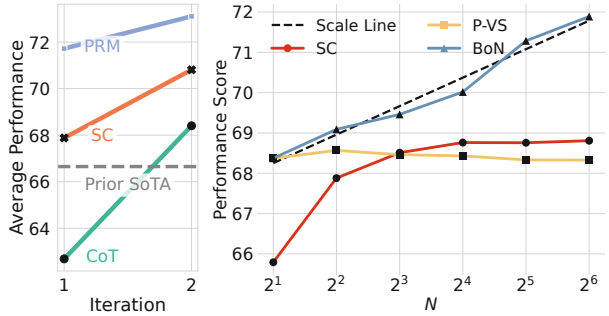


Figure 3: Interpretability evaluation for models using synthetic data, where MedS^3 produces the least hallucinatory contents among other pioneering models.



(a) (b)

Figure 4: 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.

(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 the same as (4) but initializes PRM from the SFT-tuned policy, to further show the significance of a PRM exposed to both positive and negative responses. Table 4 shows that 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 than rule-based labels, and hence outperforms previous training objectives, confirming the necessity of holistic PRM modeling.

5.2 Reliability of MedS^3

The non-eliminable hallucinations prevent most medical LLMs from being practical. Albeit inevitability, MedS^3 leverages a fine-grained soft dual-sided PRM to improve interpretability and mitigate hallucinatory contents. We leverage GPT-4o to evaluate baselines that rely on fine-tuning on synthetic datasets, including HuotuoGPT-o1, R1-Distilled-Llama-8B and UltraMedical3.1-8B, where each model's output is scored based on its medical reasonableness, logical coherence, and explainability. DrugDose is excluded from evaluation due to its small size and consequently unreliable statistical significance. Results in Fig. 3 indicates that MedS^3 achieves the highest evaluation score. We attribute such superiority to the dual awareness of the PRM, which is trained to penalize wrong intermediate steps, and therefore could induce a relatively lower score to trajectories containing hallucinations and a correct final answer. The Best-of-N strategy avoids picking up such trajectories and enhances the interpretability of the final output.

5.3 Scaling of MedS^3

In this section, we present the improvements brought by the self-evolutionary framework in Fig. 4a, and those attributable to test-time scaling in Fig. 4b. 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. 4b 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. Note that the P-VS performance is inferior to BoN, as most plau-

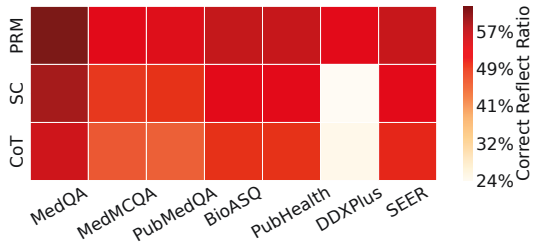


Figure 5: Reflective response ratio of MedS[3] across 7 representative datasets. Both the policy and PRM are reflection-aware to perform sequential test-time scaling.

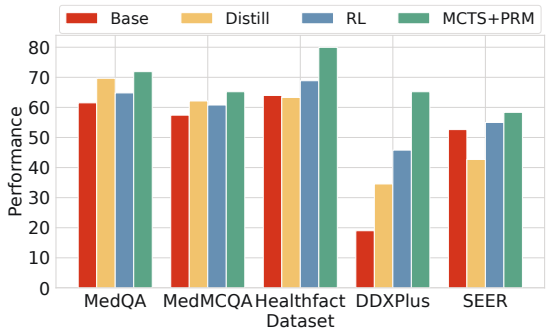


Figure 6: Three widely adopted methods to empower models with medical reasoning abilities. MCTS+PRM is the best among the three, making it the core of MedS[3].

sible reasoning chains arriving at correct answers but with incorrect reasoning steps are labeled with low values by our soft dual-sided PRM. Although these hallucinatory chains deteriorate the grouping of correct answers, our PRM still could assign the highest score for trajectories with both correct reasoning steps and final answers, therefore contributing to a log-linear scaling on BoN performance.

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. 5. 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 Reasoning Styles

In this section, we compare three reasoning 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. 2025b) 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. (2025b) to SFT the base model, and compare them with MedS[3] after the first evolution iteration. The results presented in Fig. 6 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 soundness of medical diagnosis step is clear to determine, reducing reward hacking and resulting in a more reliable PRM and credible preference estimation.

6 Related Works

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; Nori et al. 2024). Based on these, previous works develop the slow-thinking medical LLMs with distillation: Huang et al. (2025b) directly learn the reasoning trajectory generated by o1 and Chen et al. (2025) 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).

Self-Evolving Reasoning and Process Supervision Recent work in self-improving reasoning systems has explored Monte Carlo Tree Search (MCTS) (Zhang et al. 2024a) and reinforcement learning to enable models to refine their own outputs (Guo et al. 2025). Methods like Tree of Thoughts (ToT) (Yao et al. 2023) demonstrate the potential of search-based exploration for generating high-quality reasoning trajectories. Concurrently, process reward models (PRMs) have been proposed to provide step-wise feedback (Lightman et al. 2023), yet most assume binary correctness based on potential correctness and fail to penalize reasoning degradation.

7 Conclusion

In this paper, we present MedS[3], a self-evolved slow-thinking system built for universal clinical usage. We extend the clinical reasoning to diverse tasks to enhance generalization, and use MCTS 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 MedS[3] achieves superior performance on diverse medical benchmarks, especially in realistic clinical ones, surpassing open-sourced models by a large margin with fewer parameters.

Acknowledgments

This work is supported by National Key R&D Program of China (No. 2022ZD0162101), National Natural Science Foundation of China (No. 62576209) and STCSM (No. 2025SHZDZX025G05).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ankit Pal, M. S. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. website.

Biswas, A.; and Talukdar, W. 2024. Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. *arXiv preprint arXiv:2405.18346*.

Cabral, S.; Restrepo, D.; Kanjee, Z.; Wilson, P.; Crowe, B.; Abdulnour, R.-E.; and Rodman, A. 2024. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA internal medicine*, 184(5): 581–583.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; and Wang, B. 2025. Towards Medical Complex Reasoning with LLMs through Medical Verifiable Problems. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 14552–14573. Vienna, Austria: Association for Computational Linguistics.

Christophe, C.; Kanithi, P. K.; Raha, T.; Khan, S.; and Pimentel, M. A. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Dubey, S.; Tiwari, G.; Singh, S.; Goldberg, S.; and Pinsky, E. 2023. Using machine learning for healthcare treatment planning. *Frontiers in Artificial Intelligence*, 6: 1124182.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. In *Forty-second International Conference on Machine Learning*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081): 633–638.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Huang, X.; Wu, J.; Liu, H.; Tang, X.; and Zhou, Y. 2025a. m1: Unleash the Potential of Test-Time Scaling for Medical Reasoning in Large Language Models. In *The 2nd Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Huang, Z.; Geng, G.; Hua, S.; Huang, Z.; Zou, H.; Zhang, S.; Liu, P.; and Zhang, Z. 2025b. O1 Replication Journey – Part 3: Inference-time Scaling for Medical Reasoning. *arXiv preprint arXiv:2501.06458*.

Huynh, T.; He, Y.; Willis, A.; and Rueger, S. 2016. Adverse Drug Reaction Classification With Deep Neural Networks. In Matsumoto, Y.; and Prasad, R., eds., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 877–887. Osaka, Japan: The COLING 2016 Organizing Committee.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577. Hong Kong, China: Association for Computational Linguistics.

Jung, H.; Kim, Y.; Choi, H.; Seo, H.; Kim, M.; Han, J.; Kee, G.; Park, S.; Ko, S.; Kim, B.; et al. 2024. Enhancing Clinical Efficiency through LLM: Discharge Note Generation for Cardiac Patients. *arXiv preprint arXiv:2404.05144*.

Khandekar, N.; Jin, Q.; Xiong, G.; Dunn, S.; Applebaum, S.; Anwar, Z.; Sarfo-Gyamfi, M.; Safranek, C.; Anwar, A.; Zhang, A.; et al. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37: 84730–84745.

Kotonya, N.; and Toni, F. 2020. Explainable Automated Fact-Checking for Public Health Claims. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7740–7754. Online: Association for Computational Linguistics.

Lab, D. 2020. covid19-classification: Document Classification on COVID-19 Literature using the LitCovid collection and the Hedwig library. https://github.com/dki-lab/covid19-classification.

Liao, Y.; Jiang, S.; Chen, Z.; Wang, Y.; and Wang, Y. 2024. MedCare: advancing medical LLMs through decoupling clinical alignment and knowledge aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10562–10581.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Nori, H.; Usuyama, N.; King, N.; McKinney, S. M.; Fernandes, X.; Zhang, S.; and Horvitz, E. 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.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Pappas, D.; Stavropoulos, P.; Androutsopoulos, I.; and McDonald, R. 2020. BioMRC: A Dataset for Biomedical Machine Reading Comprehension. In Demner-Fushman, D.; Cohen, K. B.; Ananiadou, S.; and Tsujii, J., eds., *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 140–149. Online: Association for Computational Linguistics.

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

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Tchango, A. F.; Goel, R.; Wen, Z.; Martel, J.; and Ghosn, J. 2022. DDXPlus Dataset.

Tordjman, M.; Liu, Z.; Yuce, M.; Fauveau, V.; Mei, Y.; Hadjadj, J.; Bolger, I.; Almansour, H.; Horst, C.; Parihar, A. S.; et al. 2025. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nature medicine*, 1–1.

Tsatsaronis, G.; Schroeder, M.; Paliouras, G.; Almirantis, Y.; Androutsopoulos, I.; Gaussier, E.; Gallinari, P.; Artieres, T.; Alvers, M. R.; Zschunke, M.; et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Tu, T.; Schaekermann, M.; Palepu, A.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Cheng, Y.; et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 1–9.

Wang, P.; Li, L.; Shao, Z.; Xu, R.; Dai, D.; Li, Y.; Chen, D.; Wu, Y.; and Sui, Z. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9426–9439. Bangkok, Thailand: Association for Computational Linguistics.

Wang, T.; Jiang, Z.; He, Z.; Yang, W.; Zheng, Y.; Li, Z.; He, Z.; Tong, S.; and Gong, H. 2025. Towards Hierarchical Multi-Step Reward Models for Enhanced Reasoning in Large Language Models. *arXiv preprint arXiv:2503.13551*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Winands, M. H.; Björnsson, Y.; and Saito, J.-T. 2008. Monte-Carlo tree search solver. In *Computers and Games: 6th International Conference, CG 2008, Beijing, China, September 29-October 1, 2008. Proceedings 6*, 25–36. Springer.

Xie, Y.; Wu, J.; Tu, H.; Yang, S.; Zhao, B.; Zong, Y.; Jin, Q.; Xie, C.; and Zhou, Y. 2024. A Preliminary Study of o1 in Medicine: Are We Closer to an AI Doctor? *arXiv preprint arXiv:2409.15277*.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yonsei. 2024. MedLLAMA3-v20. website.

Yu, H.; Cheng, T.; Cheng, Y.; and Feng, R. 2025. FineMedLM-o1: Enhancing the Medical Reasoning Ability of LLM from Supervised Fine-Tuning to Test-Time Training. *arXiv preprint arXiv:2501.09213*.

Zhang, D.; Zhoubian, S.; Hu, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37: 64735–64772.

Zhang, K.; Zeng, S.; Hua, E.; Ding, N.; Chen, Z.-R.; Ma, Z.; Li, H.; Cui, G.; Qi, B.; Zhu, X.; Lv, X.; Jinfang, H.; Liu, Z.; and Zhou, B. 2024b. UltraMedical: Building Specialized Generalists in Biomedicine. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhang, Z.; Zheng, C.; Wu, Y.; Zhang, B.; Lin, R.; Yu, B.; Liu, D.; Zhou, J.; and Lin, J. 2025. The Lessons of Developing Process Reward Models in Mathematical Reasoning. *arXiv preprint arXiv:2501.07301*.

Zhao, Z.; Jin, Q.; Chen, F.; Peng, T.; and Yu, S. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10.

Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.-R.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. In *Forty-second International Conference on Machine Learning*.