Beyond Distillation: Pushing the Limits of Medical LLM Reasoning with Minimalist Rule-Based RL

Che Liu[1]*, **Haozhe Wang**[2*], **Jiazhen Pan**[3*],
Zhongwei Wan[4] , **Yong Dai**[5] , **Fangzhen Lin**[2] ,
Wenjia Bai[1] , **Daniel Rueckert**[1,3] , **Rossella Arcucci**[1]
[1]Imperial College London, [2]HKUST, [3]Technical University of Munich ,
[4]Ohio State University, [5]Fudan University
✉ *che.liu21@imperial.ac.uk*

Project page: `https://cheliu-computation.github.io/AlphaMed/`

Abstract

Improving performance on complex tasks and enabling interpretable decision making in large language models (LLMs), especially for clinical applications, requires effective reasoning. Yet this remains challenging without supervised fine-tuning (SFT) on costly chain-of-thought (CoT) data distilled from closed-source models (e.g., GPT-4o). In this work, we present **AlphaMed**, the first medical LLM to show that reasoning capability can emerge purely through reinforcement learning (RL), using minimalist rule-based rewards on public multiple-choice QA datasets, without relying on SFT or distilled CoT data. AlphaMed achieves state-of-the-art results on six medical QA benchmarks, outperforming models trained with conventional SFT+RL pipelines. On challenging benchmarks (e.g., MedXpert), AlphaMed even surpasses larger or closed-source models such as DeepSeek-V3-671B and Claude-3.5-Sonnet. To understand the factors behind this success, we conduct a comprehensive data-centric analysis guided by three questions: **(i)** Can minimalist rule-based RL incentivize reasoning without distilled CoT supervision? **(ii)** How do dataset quantity and diversity impact reasoning? **(iii)** How does question difficulty shape the emergence and generalization of reasoning? Our findings show that dataset informativeness is a key driver of reasoning performance, and that minimalist RL on informative, multiple-choice QA data is effective at inducing reasoning without CoT supervision. We also observe divergent trends across benchmarks, underscoring limitations in current evaluation and the need for more challenging, reasoning-oriented medical QA benchmarks. The code and pretrained model weights will be publicly released upon acceptance.

1 Introduction

Recently, the reasoning capabilities of large language models (LLMs) have advanced significantly, achieving impressive results in tasks requiring complex reasoning, such as mathematical problem solving, code generation, and general-purpose benchmarks [1–4]. These developments highlight the potential of LLMs to generalize and perform multi-step reasoning across domains. In the medical domain, reasoning is particularly crucial. Clinical natural language processing (NLP) tasks often require interpreting nuanced patient information, integrating knowledge from diverse sources, and making informed decisions [5–7]. More importantly, reasoning provides a valuable lens into the

*Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

model's decision-making process, allowing researchers and clinicians to examine how conclusions are derived. This improves the interpretability and transparency of AI outputs, which are essential for clinical trust [8, 9].

Currently, most medical LLMs acquire reasoning capabilities through supervised fine-tuning (SFT) on chain-of-thought (CoT) datasets, often followed by reinforcement learning (RL) for further refinement. However, this pipeline heavily relies on an initial SFT stage using costly CoT data, which are either manually crafted or distilled from closed-source commercial models such as GPT-4o [10, 11]. This dependence not only incurs substantial annotation and distillation costs but also introduces scalability and accessibility challenges, as it ties model development to expensive and external resources. These limitations motivate a critical question:

Can we achieve medical reasoning through minimalist rule-based RL
without relying on distilled CoT data?

To address this question, we propose **AlphaMed**, the first work designed to incentivize reasoning capability solely through minimalist rule-based RL, going beyond conventional approaches that rely on SFT with CoT data. Instead of depending on distilled CoT data supervision, AlphaMed is trained directly via simple rule-based rewards derived from multiple-choice QA datasets. Our key contributions are as follows:

- We show that minimalist rule-based RL can incentivize reasoning ability in medical LLMs without relying on distilled CoT data, achieving superior performance. We further analyze how dataset quantity, diversity, and especially informativeness impact reasoning performance. We empirically find that higher informativeness enhances reasoning performance, while less-informative data limits gains.
- We show that reasoning can be incentivized even with lower-difficulty data and further enhanced by harder examples. While high-difficulty samples benefit challenging benchmarks like MedXpert, a mix of difficulty levels is essential for robust generalization. Non-monotonic trends across benchmarks suggest that current evaluations may be insufficient to assess medical LLM reasoning.
- Building on these insights, we introduce AlphaMed, a medical LLM trained solely via minimalist rule-based RL without any SFT on distilled CoT data, and demonstrate that it achieves state-of-the-art performance across six mainstream medical QA benchmarks, outperforming models that use complex training strategies with CoT data and even surpassing larger or closed-source models such as DeepSeek-V3-671B and GPT-4o.

2 Related Work

Supervised Fine-Tuning for Reasoning in LLMs. Large language models can acquire complex reasoning skills through SFT on CoT data. For example, [12] showed that training models to generate step-by-step reasoning paths significantly improves performance on math and logic problems. [13] scaled this approach by incorporating a broad range of CoT examples into instruction tuning across diverse tasks. [14] proposed STaR, where a model bootstraps its own reasoning traces to reduce reliance on human-annotated CoT. However, recent work [15] suggests that SFT often encourages memorization of training rationales rather than true reasoning generalization, limiting robustness in out-of-distribution or unfamiliar tasks. Moreover, obtaining high-quality CoT data is costly, requiring either expert annotations or distillation from proprietary models, posing significant challenges to scalability and adaptability [16, 17].

Reinforcement Learning with Preference Data after SFT. InstructGPT [18] introduced reinforcement learning with human preferences (RLHF) to align model behavior with user intent. Subsequent research has shown that RL can enhance generalization [16, 19] and better capture nuanced human preferences beyond rote memorization [15]. Among RL algorithms, Proximal Policy Optimization (PPO) is widely used, but it is highly resource-intensive—requiring learned reward models that are often sensitive to noise, difficult to interpret, and occasionally misaligned with intended objectives [20]. To address these limitations, Direct Preference Optimization (DPO) [20] eliminates the need for an explicit reward model by directly optimizing over preference pairs. However, DPO

still relies on high-quality preference annotations, which are particularly challenging to construct in the medical domain due to clinical ambiguity and a lack of universal agreement on what constitutes a "better" response [21]. Recently, DeepSeek-R1-Zero [22] demonstrated that reasoning behavior can be effectively elicited without CoT supervision or preference annotations, instead by leveraging final answers (e.g., multiple-choice accuracy) as rule-based supervision signals [16, 19, 23, 24].

Open-Source Medical LLMs. Open-source medical LLMs have emerged as promising tools for domain-specific clinical reasoning, yet most remain heavily dependent on supervised data or handcrafted feedback. HuatuoGPT [25] was instruction-tuned on ChatGPT-distilled medical dialogues. BioMistral [26] adapted the Mistral architecture to biomedical question answering through continued pretraining [27] and domain-specific instruction tuning. OpenBioLLM [21] and UltraMedical [28] utilized DPO-based preference optimization, but their preference pairs were directly distilled from closed-source models, making supervision ambiguous and potentially inconsistent with expert clinical reasoning. Since human verification of each distilled example is prohibitively costly and impractical, there is no guarantee that the reasoning process reflected in the supervision is valid. HuatuoGPT-o1 [29] further incorporated PPO using a self-trained 3B reward model and relied on CoT data distilled from OpenAI o1. However, this approach is resource-intensive and tightly coupled to the quality and coverage of proprietary data, limiting its scalability and generalizability. m1 [30] also adopts SFT on distilled chain-of-thought data, where step-by-step reasoning traces are generated by external large reasoning model, thus still relying on distilled CoT data.

3 Preliminaries

Group Relative Policy Optimization (GRPO) Given a question-answer pair (q, a), the behaviour policy π_{old} generates a set of G candidate completions $\{o_i\}_{i=1}^{G}$ for each question q. Each response receives a scalar reward r_i, which may be derived from human preference comparisons or automated scoring heuristics; in this work, we use a rule-based reward. The relative quality of each response is assessed within the group through normalization. The training objective is:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big(\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\Big)\Big)\Bigg] \tag{1}$$

where the group-normalized advantage $\hat{A}_{i,t}$ and the token-level importance weight $r_{i,t}(\theta)$ are defined as:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}, \quad r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_i, < t)}{\pi_{\text{old}}(o_{i,t} \mid q, o_i, < t)}.$$

Here, ϵ is a hyperparameter controlling the tolerance for policy deviation. The `clip` function prevents large updates by ensuring that the ratio between the current and reference policy stays within a predefined range. Specifically, it clips the importance weight $r_{i,t}(\theta)$ to the interval $[1-\epsilon, 1+\epsilon]$, thereby stabilizing training and mitigating the risk of policy collapse. This objective encourages the model to improve token probabilities for completions with above-average rewards, while stabilizing updates via a clipped importance weight similar to PPO [31].

Rule-based Reward Modelling To enable minimalist RL without relying on external verifiers or human-provided rewards, we adopt a simple rule-based approach consistent with [22]. This method directly evaluates the correctness of the model's output using binary feedback, eliminating the need for a separate reward model:

$$r_i = \begin{cases} 1, & \text{if } \texttt{is\_answer\_correct}(\hat{y}_i, y) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Here, y is the ground-truth answer and $\hat{y}_i$ denotes the model-generated prediction from the i-th output o_i. This straightforward reward mechanism provides a clear supervision signal grounded in task accuracy. By leveraging structured outputs (e.g., multiple-choice answers), we enable effective RL without manually written rationales or preference annotations.

4 AlphaMed

4.1 Training Configuration

We aim to elicit medical reasoning behavior purely through rule-based RL, without relying on SFT with CoT data or RL with rewards from external verifiers. To ensure a fair comparison with `HuatuoGPT-o1` [29], we adopt `Llama3.1-8B-Instruct` and `Llama3.1-70B-Instruct` as backbone models. All experiments are conducted under full parameter tuning with a batch size of 512, meaning each batch contains 64 QA pairs and each question generates 8 candidate answers, trained for 300 steps. We use `verl`[2] [32], a framework designed for rule-based RL. A simple binary reward function, defined in Eq. 2, assigns 1 if the model's response ends with a correctly formatted boxed answer matching the ground truth (e.g., `\boxed{C}`), and 0 otherwise. The model is optimized using the GRPO objective described in Eq. 1. We train the 8B model on 8 Nvidia A800-80G GPUs and the 70B model on 64 A800-80G GPUs.

4.2 Evaluation Configuration

Datasets. We evaluate our models on six medical QA benchmarks, using accuracy as the evaluation metric across all datasets. These include MedQA-USMLE [33] (MedQA), MedMCQA [34] (MedMCQA), PubMedQA [35] (PubMedQA), MMLU-Pro medical subsets [36] (MMLU-ProM), GPQA medical subsets [37] (GPQA-M), and the most recent and challenging large-scale dataset, MedXpertQA [38] (MedXpert). Details are provided in Sec. A.2.

Based on their levels of challenge [39], we categorize MedQA, MedMCQA, and PubMedQA [33–35] as ***normal***, while MMLU-ProM and GPQA-M [40, 37] are classified as ***hard***, as they primarily target advanced expert-level knowledge. Finally, MedXpert [38] is designated as ***hard+***, as the original work explicitly highlights its focus on complex clinical reasoning and expert-level decision making, positioning it as one of the most challenging benchmarks to date.

Baseline Methods. We compare against a broad range of general and medical-specific LLM baselines. General-purpose base instruct models include `Qwen2.5-7B/32B/72B` and `Llama3.1-8B/70B`. Medical-specific models cover `MedLlama3`, `OpenBioLLM` [41], `MMed` and `MMed-S` [42], `Med42` [43], and `UltraMedical` [28], which leverage distilled preference data and RL following SFT. `HuatuoGPT-o1` [29] is trained on CoT data distilled from GPT-4o using model-based RL with a large (3B) reward model. `m1` [30] is similarly trained with extensive CoT distilled from DeepSeekR1 [22] via SFT.

5 Experiments

5.1 Data Curation

Initial Data Collection. Following [30], we collect the training splits of three large-scale public multiple-choice medical QA datasets: MedQA [44], MedMCQA [45], and PubMedQA [35][3][4]. MedQA [44] contains expert-level clinical questions from the USMLE. MedMCQA [45] includes factoid and reasoning questions from Indian medical entrance exams (AIIMS, NEET). PubMedQA [35] focuses on biomedical research question answering. Notably, its training split is automatically generated by a machine learning model that heuristically converts biomedical research article abstract into yes/no questions and assigns answers based on negation cues. The dataset statistics are summarized in Sec. A.1.

Quantifying Data Difficulty. To quantify question difficulty, we perform inference using `Llama3.1-8B-Instruct` [46]. For each question, we generate five reasoning completions with the following prompt: "`Please reason step by step, and put the final answer in \boxed{}`". We then calculate the proportion of correct predictions among the five outputs, which serves as a proxy for the question's difficulty. Based on this proportion, we categorize questions into six difficulty levels (L1–L6). Specifically, L1 includes questions where all five comple-

[2] https://github.com/volcengine/verl

[3] We use the official training splits of all three datasets.

[4] For PubMedQA [35], only questions with definitive answer labels (i.e., A/B/C) are retained.

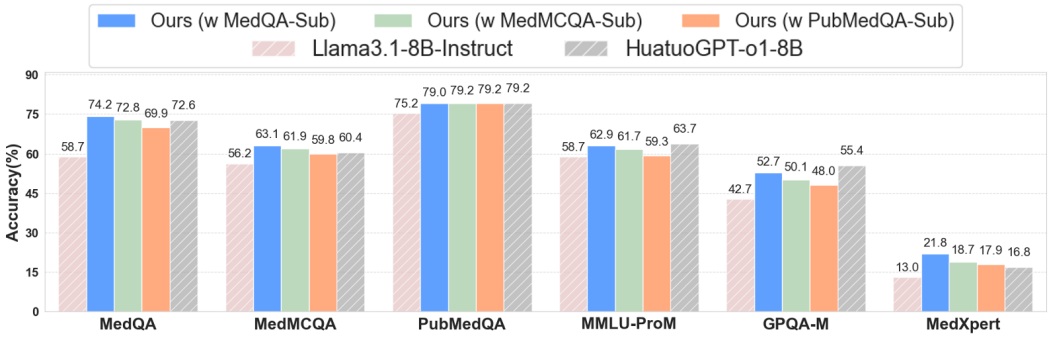

Figure 1: **Performance comparison on six medical QA benchmarks.** Our models are initialized with `Llama3.1-8B-Instruct` [46] and trained using minimalist rule-based RL on one of three balanced subsets: *MedQA-Sub*, *MedMCQA-Sub*, or *PubMedQA-Sub* (shown as blue, green, and orange bars, respectively). Despite using only 1,200 examples per subset, all variants of our model achieve substantial improvements over the base `Llama3.1-8B-Instruct` and match or surpass the strong baseline `HuatuoGPT-o1-8B` across all benchmarks.

tions are correct, L2 where four are correct, and so on, with L6 representing questions where all five completions are incorrect. The difficulty level distribution of each train set as shown in Tab. 2

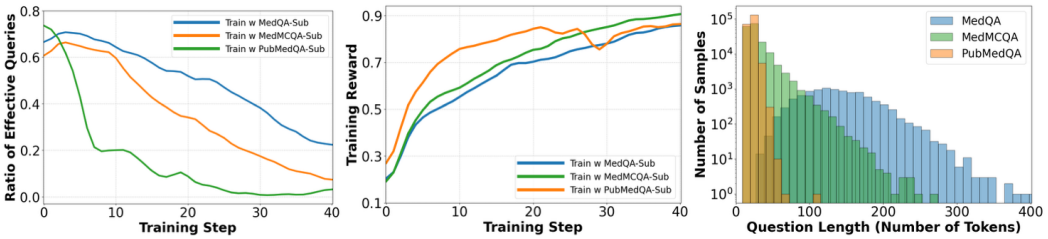

Figure 2: **Dataset analysis and training dynamics. Left:** Ratio of effective queries over training steps; each curve corresponds to models trained on a specific subset. **Middle:** Training reward per step for models trained on each subset. **Right:** Distribution of question lengths (number of tokens) in MedQA, MedMCQA, and PubMedQA [44, 45, 35].

5.2 RQ1: Can Minimalist RL Incentivize Medical Reasoning Without Distilled-CoT SFT?

To investigate whether minimalist rule-based RL can incentivize medical reasoning in LLMs without relying on SFT with distilled CoT data, we conduct a pilot study by sampling 200 examples from each difficulty level to construct three balanced subsets (1,200 samples each) from three public medical QA datasets: *MedQA-Sub*, *MedMCQA-Sub*, and *PubMedQA-Sub*. We use `Llama3.1-8B-Instruct` as the backbone model and train it separately on each subset using minimalist RL. As shown in Fig. 1, all models trained on these subsets achieve substantial gains over the original backbone across all six benchmarks (e.g., +15.5% on MedQA, +8.8% on MedXpert). Remarkably, all variants trained on different subsets perform comparably to or even surpass `HuatuoGPT-o1-8B` [47], a strong baseline trained via SFT on CoT data distilled from GPT-4o [48] and further fine-tuned with RL using a 3B reward model. Notably, on MedXpert [38], the most challenging benchmark, all three variants outperform `HuatuoGPT-o1-8B` [47]. These results demonstrate that reasoning capability can be effectively incentivized through minimalist RL on small-scale, low-cost multiple-choice QA data, without relying on SFT with distilled CoT data, and can even outperform models trained with more complex strategies.

Surprisingly, **multistep reasoning** (e.g., `Step 1..., Step 2...`; see Fig. 11, 12, 13) spontaneously emerges in the model's output, which derives the final answer through sequential analysis, despite being supervised only on the final choice, without intermediate reasoning traces like distilled CoT data [30, 47]. This emergent behavior shows that minimalist rule-based RL not only boosts performance but also encourages structured reasoning, offering valuable interpretability into the model's decision-making.

Performance Variation and Training Dynamics Across Subsets. We observe clear performance differences among training subsets, consistently ranking as *MedQA-Sub* > *MedMCQA-Sub* >

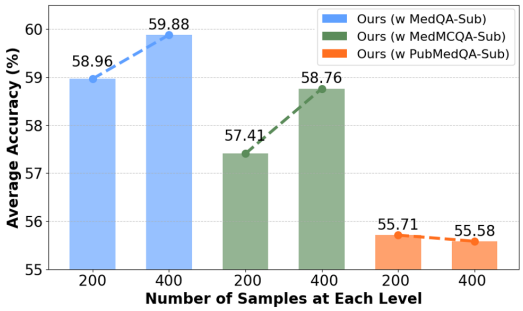

Figure 3: **Effect of data quantity.** Average accuracy across six medical QA benchmarks as the number of samples per level increases from 200 to 400, resulting in the total subset size growing from 1,200 to 2,400 examples. Scaling *MedQA-Sub* and *MedMCQA-Sub* leads to consistent performance gains, highlighting the value of informative data. In contrast, *PubMedQA-Sub* shows no improvement, reflecting the limitations of low-informative data sources.

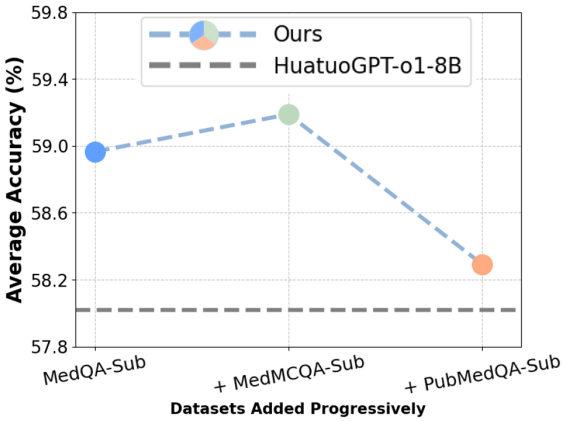

Figure 4: **Effect of data diversity.** Average accuracy across six medical QA benchmarks when models are trained individually on single or combined subsets. Adding *MedMCQA-Sub* to *MedQA-Sub* boosts performance, while further adding *PubMedQA-Sub* reduces it, suggesting that less informative data can negate the benefits of increased diversity.

PubMedQA-Sub. To understand this variation, we explore the training dynamics of models trained on each subset. As depicted in Fig. 2 (left), following [16], the ratio of effective queries is computed as $1 - \frac{\#\text{solved all}+\#\text{solved none}}{\#\text{unique queries}}$, where "solved all" and "solved none" denote batches in which all responses are either correct or incorrect. Models trained on *PubMedQA-Sub* exhibit a rapid decline in the effective query ratio, indicating premature saturation and a reduction in effective samples from the batch. The training reward in Fig. 2 (middle) further supports this: the *PubMedQA-Sub* variant starts with a higher initial reward and increases rapidly, suggesting that the data is easy to learn at the start, but quickly saturates after about 20 steps. In contrast, the *MedQA-Sub* and *MedMCQA-Sub* models improve steadily throughout training.

Dataset Informativeness as a Key Driver. To further investigate these dynamics, we analyze the question length distributions in the source datasets of each subset, as shown in Fig. 2 (right). Notably, MedQA [44] exhibits a significantly longer question length distribution compared to MedMCQA [45] and PubMedQA [35], this ordering closely matches the observed performance of model variants trained on the respective subsets. These differences are linked to dataset construction mechanisms: PubMedQA [35] is automatically curated from biomedical literature, often resulting in noisier and less informative questions; MedMCQA [45] is based on human-authored medical school entrance exams, providing more reliable and informative samples; MedQA [44] is sourced from the USMLE, a challenging licensing exam, and thus contains the most informative and well-structured questions. Altogether, our findings suggest that question length serves as a practical proxy for dataset informativeness in medical QA. High-informativeness, exam-certified data provide more stable and effective learning signals for minimalist RL, whereas noisy, automatically curated data may offer lower informativeness and thus hinder the acquisition of reasoning ability.

Finding 1.1: Minimalist rule-based RL enables medical reasoning in LLM beyond reliance on SFT with distilled CoT data.

Finding 1.2: Dataset informativeness is critical for training success. LLM trained on low informative or noisy data exhibit degraded performance. Question length serves as a practical proxy for informativeness in medical QA.

5.3 RQ2: Impact of Dataset Quantity and Diversity

Effect of Dataset Quantity. To investigate the effect of training data size, we increase the number of samples per difficulty level from 200 to 400 for each of the three subsets, resulting in the total

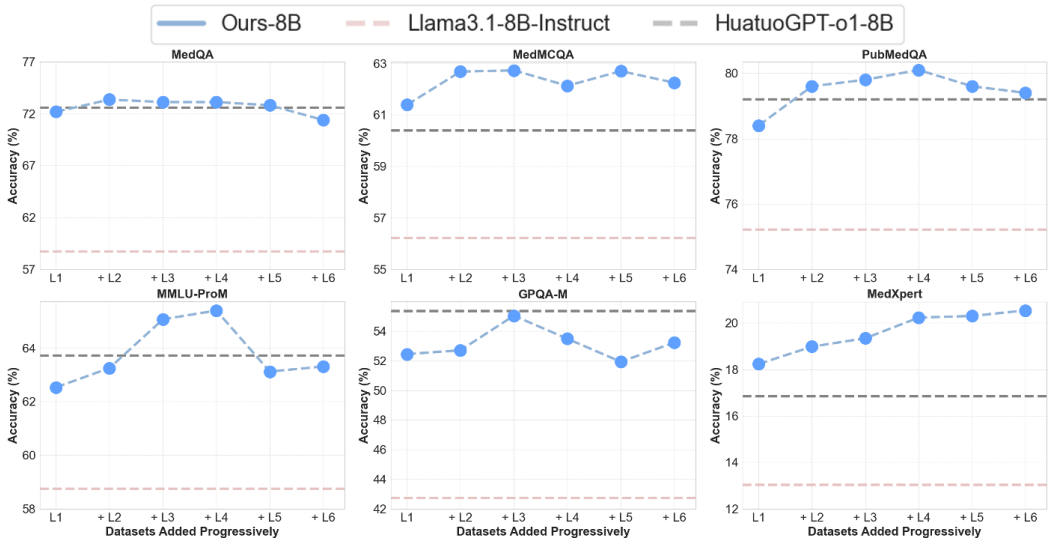

Figure 5: Performance on six benchmarks when training on subsets with increasing difficulty levels (L1 to L6). Each blue dot represents a separately trained model on a subset that includes all data up to the indicated difficulty level; new data are incorporated only through separate training runs, not incrementally during training. While performance on MedXpert [38] increases consistently, trends on other benchmarks vary. Final models trained on the full set (L1–L6) generally achieve comparable or superior performance to HuatuoGPT-o1-8B [47].

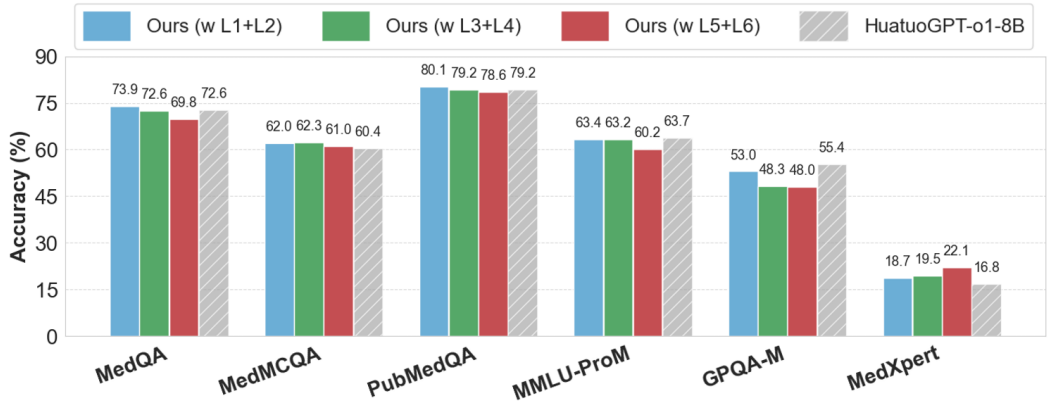

Figure 6: Performance on six benchmarks when training with distinct difficulty groups: easy (L1+L2), medium (L3+L4), and hard (L5+L6). While harder training data improves MedXpert [38] accuracy, performance on other benchmarks declines, suggesting that relying solely on difficult samples may impair general reasoning ability.

number of samples in each subset increasing from 1,200 to 2,400. As shown in Fig. 3, we report the average accuracy across six benchmarks. Scaling *MedQA-Sub* improves accuracy from 58.96% to 59.88%, and *MedMCQA-Sub* improves from 57.41% to 58.76%, demonstrating that increasing high-informative data benefits model performance. In contrast, scaling *PubMedQA-Sub* yields no improvement (55.71% $\rightarrow$ 55.58%), suggesting that adding more low-informative or noisy samples may degrade performance rather than enhance it.

Effect of Dataset Diversity. We further examine the effect of dataset diversity by progressively combining subsets. As shown in Fig. 4, adding *MedMCQA-Sub* to *MedQA-Sub* further improves performance, highlighting the benefit of combining diverse and informative datasets. However, incorporating *PubMedQA-Sub* reverses the upward trend and leads to a decline in performance, indicating that noisy and less informative data not only fail to contribute but may also harm reasoning ability.

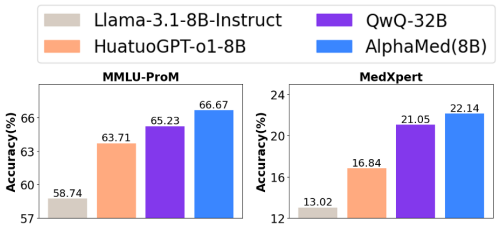

Figure 7: Comparison of AlphaMed(8B) with prior models on MMLU-ProM [36] and MedXpert [38]. Despite its smaller scale and use of minimalist RL, AlphaMed(8B) outperforms the larger model QwQ-32B [49] and other baselines.

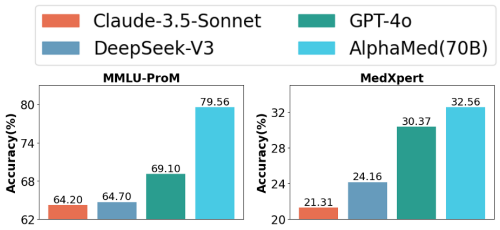

Figure 8: **AlphaMed(70B)** achieves superior performance over Claude-3.5-Sonnet [50], GPT-4o [48], and DeepSeek-V3 (671B) [51] on MMLU-ProM [36] and MedXpert [38], showcasing its strong reasoning ability.

Finding 2: Performance improves with increased data quantity and diversity only when the additional samples are informative; low-quality data harms the learning of reasoning ability.

5.4 RQ3: Impact of Dataset Quality

We analyze how increasing training difficulty affects performance across six benchmarks, as shown in Fig. 5. MedQA, MedMCQA, and PubMedQA [44, 45, 35] exhibit inverse U-shaped trends, performance peaks with moderate difficulty (L1–L4) and declines with harder samples (L5–L6), suggesting diminishing returns from high-difficulty data. In contrast, MMLU-ProM [40] and GPQA-M [37] show oscillating patterns, while MedXpert [38] improves steadily with increasing difficulty, highlighting the value of harder samples for complex tasks. To validate this, we train models on three difficulty groups (easy: L1+L2, medium: L3+L4, hard: L5+L6; Fig. 6). On MedXpert [38], models trained on hard samples perform best, confirming their role in promoting advanced reasoning. For other benchmarks, training on easy and medium levels yields better generalization, while hard-only training underperforms.

Emerging Reasoning Capability from Simple Data, Indicating Benchmark Limits. Interestingly, models trained only on L1+L2 (a total of 2,400 samples) already match or surpass HuatuoGPT-o1-8B [47] on several benchmarks. As shown in Fig. 5, even on MedXpert, only training with L1 data exceeds HuatuoGPT-o1-8B [47], with further gains from adding more levels, indicating that reasoning can emerge from simple data. These findings underscore the importance of balanced training difficulty to support broad generalization. They also reveal a potential pitfall: if high benchmark scores can be achieved without exposure to difficult samples, such scores may not reflect genuine reasoning ability, raising concerns about the adequacy of current benchmark designs.

Finding 3.1: Mixed difficulty training is crucial for generalizable reasoning.

Finding 3.2: Current benchmarks may insufficient to capture true reasoning progress.

5.5 Main Results

Building on the above findings which highlight the importance of dataset quantity, diversity, informativeness, and mixed difficulty for incentivizing reasoning, we construct our final training set accordingly. Specifically, we include all samples from MedQA [44] due to its high informativeness, and sample 1,600 QA pairs from each difficulty level of MedMCQA [45] to match the overall scale of MedQA [44]. PubMedQA [35] is excluded due to its limited informativeness and the performance degradation observed when it is included, as discussed in RQ1 and RQ2. The final training set comprises 19,178 QA pairs. This dataset is used to train our final models: **AlphaMed(8B)**, based on `Llama3.1-8B-Instruct`, and **AlphaMed(70B)**, based on `Llama3.1-70B-Instruct`, both optimized using minimalist rule-based RL. Since MedQA [44] and MedMCQA [45] are used for training, we treat PubMedQA [35], MMLU-ProM [40], GPQA-M [52], and MedXpert [38] as out-of-domain (OOD) benchmarks.

We present the full results in Tab. 1. Across both model scales, `AlphaMed` consistently outperforms all compared methods on both in-domain and OOD benchmarks, using only minimalist rule-based RL and multiple-choice QA supervision. Remarkably, this advantage holds even against models trained with more complex strategies [47, 28], including SFT on distilled CoT data [47, 28, 30]

Model	MedQA	MedMCQA	PubMedQA	MMLU-ProM	GPQA-M	MedXpert
	In-Domain		*Out-of-Domain*			
Challenge Level	*Normal*	*Normal*	*Normal*	*Hard*	*Hard*	*Hard+*
< 10B LLMs						
Llama-3.1-8B-Instruct	58.72	56.21	75.21	58.74	42.73	13.02
Qwen2.5-7B-Instruct	61.51	56.56	71.30	61.17	42.56	12.15
Qwen2.5-7B-Instruct^{+}	64.49	56.11	72.60	62.15	52.56	13.18
MedLlama3-8B-v1	55.07	34.74	52.70	27.43	30.77	11.04
MedLlama3-8B-v2	59.39	59.34	75.50	55.11	36.41	13.46
MMed-8B†‡	54.28	52.71	63.40	48.27	34.87	13.73
MMedS-8B†‡	57.19	47.29	77.50	33.55	22.05	17.39
MMed-8B-EnIns†‡	60.33	58.09	63.80	51.60	45.90	18.56
Med42-8B‡	59.78	56.35	76.00	55.64	48.21	14.63
OpenBioLLM-8B$^{\dagger\ddagger\diamond}$	55.30	54.63	70.10	49.32	41.03	14.29
UltraMedical-8B-3$^{\dagger\ddagger\diamond}$	71.09	59.22	71.00	61.50	50.00	15.25
UltraMedical-8B-3.1$^{\dagger\ddagger\diamond}$	75.73	63.78	79.20	64.30	48.72	17.39
HuatuoGPT-o1-8B$^{\dagger\ddagger\diamond}$	72.60	60.40	79.20	63.71	55.38	16.84
m1-7B†‡	75.81	62.54	75.80	65.86	53.08	19.81
AlphaMed(8B)	**76.19**	**64.47**	**80.40**	**66.67**	**58.44**	**22.14**
> 10B LLMs						
Llama-3.1-70B-Instruct	78.42	72.53	78.52	74.50	55.73	21.32
QwQ-32B	78.62	69.71	77.85	65.23	56.92	21.05
Qwen2.5-32B-Instruct	75.26	64.83	68.00	74.72	63.85	13.87
Qwen2.5-32B-Instruct^{+}	74.86	64.33	68.90	74.72	64.87	14.56
Qwen2.5-72B-Instruct	74.55	66.60	70.80	66.06	62.05	14.91
Qwen2.5-72B-Instruct^{+}	76.43	66.15	71.30	69.77	63.85	19.65
Med42-70B‡	51.14	62.28	78.10	54.53	50.77	16.29
OpenBioLLM-70B$^{\dagger\ddagger\diamond}$	75.10	74.23	79.30	71.92	50.77	21.33
UltraMedical-70B-3$^{\dagger\ddagger\diamond}$	83.90	72.94	80.00	73.94	58.72	21.67
HuatuoGPT-o1-70B$^{\dagger\ddagger\diamond}$	83.30	73.60	80.60	76.09	66.67	26.36
m1-32B†‡	83.50	67.34	77.60	77.94	66.67	25.53
AlphaMed(70B)	**87.52**	**75.09**	**80.90**	**79.56**	**77.46**	**32.56**

Table 1: Combined performance of models on six medical QA benchmarks with varying levels of challenge. In-domain and out-of-domain tasks, as well as challenge levels (Normal, Hard, Hard+), are indicated below the task names. **m1** denotes models that use *test-time scaling during inference*. $^{+}$: using CoT prompting during inference; †: trained with distilled CoT data from stronger models (e.g., GPT-4o); ‡: trained with external datasets beyond MedQA and MedMCQA; $^{\diamond}$: trained via RL with verifier reward models or distilled preference data from powerful models (e.g., GPT-4o). **AlphaMed (Ours)** is trained solely with minimalist rule-based RL on multi-choice QA, without any SFT on distilled CoT data, preference data, or rewards from verifiers.

and methods enhanced with test-time scaling [30]. Notably, `AlphaMed(8B)` surpasses the larger reasoning model `QwQ-32B` [49] on challenging OOD benchmarks, as shown in Fig. 7. At the 70B scale, `AlphaMed(70B)` outperforms even closed-source models such as `GPT-4o` [48] and `Claude-3.5-Sonnet` [50], as well as the open-source `DeepSeek-V3` (671B parameters) [51], as shown in Fig. 8. These results show that minimalist rule-based RL, trained with a well-constructed multiple-choice QA dataset, enables effective and scalable medical reasoning in LLMs without relying on distilled CoT supervision.

6 Conclusion

We present AlphaMed, the first work to demonstrate that reasoning capabilities can emerge solely through minimalist rule-based RL, without relying on SFT with distilled CoT data. By leveraging only multiple-choice QA datasets, AlphaMed achieves state-of-the-art performance across six diverse and challenging medical QA benchmarks, surpassing models trained with conventional SFT+RL pipelines, and even outperforming closed-source models (e.g., GPT-4o [48]. Through comprehensive data-centric analyses, we show that reasoning ability can be effectively incentivized by selecting data based on informativeness. We further find that increasing the number of informative training samples improves performance, and that varying difficulty levels contribute differently across benchmarks, underscoring the importance of mixing difficulty to promote generalizable reasoning. A well-curated dataset with high informativeness and diverse difficulty levels is key to

advancing reasoning, without requiring handcrafted rationales or distilled data from closed models. Our findings also reveal a critical caveat: while challenging benchmarks benefit from harder training samples, others exhibit mixed or plateauing trends, suggesting that existing benchmarks may be insufficient to evaluate progress of reasoning ability. This highlights the need for more challenging, reasoning-oriented benchmarks. Altogether, AlphaMed not only establishes a strong medical LLM, but also offers insights into how models reach final predictions through emergent reasoning, encouraging further exploration of interpretable systems in medical NLP.

References

[1] Y. Qin, X. Li, H. Zou, Y. Liu, S. Xia, Z. Huang, Y. Ye, W. Yuan, H. Liu, Y. Li *et al.*, "O1 replication journey: A strategic progress report–part 1," *arXiv preprint arXiv:2410.18982*, 2024.

[2] Z. Zeng, Q. Cheng, Z. Yin, B. Wang, S. Li, Y. Zhou, Q. Guo, X. Huang, and X. Qiu, "Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective," *arXiv preprint arXiv:2412.14135*, 2024.

[3] J. Wang, M. Fang, Z. Wan, M. Wen, J. Zhu, A. Liu, Z. Gong, Y. Song, L. Chen, L. M. Ni *et al.*, "Openr: An open source framework for advanced reasoning with large language models," *arXiv preprint arXiv:2410.09671*, 2024.

[4] M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Heylar, R. Dias, A. Vallone, H. Ren, J. Wei, H. W. Chung, S. Toyer, J. Heidecke, A. Beutel, and A. Glaese, "Deliberative alignment: Reasoning enables safer language models," *OpenAI Blog*, 2024. [Online]. Available: https://openai.com/index/deliberative-alignment/

[5] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi *et al.*, "Capabilities of gemini models in medicine," *arXiv preprint arXiv:2404.18416*, 2024.

[6] J. Chen, C. Gui, A. Gao, K. Ji, X. Wang, X. Wan, and B. Wang, "Cod, towards an interpretable medical agent using chain of diagnosis," *arXiv preprint arXiv:2407.13301*, 2024.

[7] V. L. Patel, J. F. Arocha, and J. Zhang, "Thinking and reasoning in medicine," *The Cambridge handbook of thinking and reasoning*, vol. 14, pp. 727–750, 2005.

[8] S. Xu, Y. Zhou, Z. Liu, Z. Wu, T. Zhong, H. Zhao, Y. Li, H. Jiang, Y. Pan, J. Chen *et al.*, "Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios," *arXiv preprint arXiv:2411.14461*, 2024.

[9] M.-H. Temsah, A. Jamal, K. Alhasan, A. A. Temsah, and K. H. Malki, "Openai o1-preview vs. chatgpt in healthcare: A new frontier in medical ai reasoning," *Cureus*, vol. 16, no. 10, p. e70640, 2024.

[10] Y. Xie, J. Wu, H. Tu, S. Yang, B. Zhao, Y. Zong, Q. Jin, C. Xie, and Y. Zhou, "A preliminary study of o1 in medicine: Are we closer to an ai doctor?" *arXiv preprint arXiv:2409.15277*, 2024.

[11] J. Chen, X. Wang, K. Ji, A. Gao, F. Jiang, S. Chen, H. Zhang, D. Song, W. Xie, C. Kong *et al.*, "Huatuogpt-ii, one-stage training for medical adaption of llms," *arXiv preprint arXiv:2311.09774*, 2023.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[13] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, J. Wei *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[14] E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman, "Star: Bootstrapping reasoning with reasoning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 476–15 488, 2022.

[15] T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma, "Sft memorizes, rl generalizes: A comparative study of foundation model post-training," *arXiv preprint arXiv:2501.17161*, 2025.

[16] H. Wang, C. Qu, Z. Huang, W. Chu, F. Lin, and W. Chen, "Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning," *arXiv preprint arXiv:2504.08837*, 2025.

[17] A. Su, H. Wang, W. Ren, F. Lin, and W. Chen, "Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning," *arXiv preprint arXiv:2505.15966*, 2025.

[18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[19] H. Wang, L. Li, C. Qu, F. Zhu, W. Xu, W. Chu, and F. Lin, "To code or not to code? adaptive tool integration for math language models via expectation-maximization," *arXiv preprint arXiv:2502.00691*, 2025.

[20] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[21] A. Ura, "Openbiollm-70b: Advancing open-source biomedical llms with direct preference optimization," *Hugging Face Blog*, 2024, available at https://huggingface.co/blog/aaditya/openbiollm.

[22] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[23] H. Zeng, D. Jiang, H. Wang, P. Nie, X. Chen, and W. Chen, "Acecoder: Acing coder rl via automated test-case synthesis," *arXiv preprint arXiv:2502.01718*, 2025.

[24] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, and D. Rueckert, "Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning," *arXiv preprint arXiv:2502.19634*, 2025.

[25] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, J. Li, G. Chen, X. Wu, Z. Zhang, Q. Xiao *et al.*, "Huatuogpt, towards taming language model to be a doctor," *arXiv preprint arXiv:2305.15075*, 2023.

[26] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pretrained large language models for medical domains," *arXiv preprint arXiv:2402.10373*, 2024.

[27] Z. Ke, Y. Shao, H. Lin, T. Konishi, G. Kim, and B. Liu, "Continual pre-training of language models," *arXiv preprint arXiv:2302.03241*, 2023.

[28] K. Zhang, S. Zeng, E. Hua, N. Ding, Z.-R. Chen, Z. Ma, H. Li, G. Cui, B. Qi, X. Zhu *et al.*, "Ultramedical: Building specialized generalists in biomedicine," *Advances in Neural Information Processing Systems*, vol. 37, pp. 26 045–26 081, 2024.

[29] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "Huatuogpt-o1: Towards medical complex reasoning with llms," *arXiv preprint arXiv:2412.18925*, 2024.

[30] X. Huang, J. Wu, H. Liu, X. Tang, and Y. Zhou, "m1: Unleash the potential of test-time scaling for medical reasoning with large language models," *arXiv preprint arXiv:2504.00869*, 2025.

[31] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[32] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu, "Hybridflow: A flexible and efficient rlhf framework," *arXiv preprint arXiv:2409.19256*, 2024.

[33] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

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

[35] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[36] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[37] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q&a benchmark," in *First Conference on Language Modeling*, 2024.

[38] Y. Zuo, S. Qu, Y. Li, Z. Chen, X. Zhu, E. Hua, K. Zhang, N. Ding, and B. Zhou, "Medxpertqa: Benchmarking expert-level medical reasoning and understanding," *arXiv preprint arXiv:2501.18362*, 2025.

[39] X. Tang, D. Shao, J. Sohn, J. Chen, J. Zhang, J. Xiang, F. Wu, Y. Zhao, C. Wu, W. Shi *et al.*, "Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning," *arXiv preprint arXiv:2503.07459*, 2025.

[40] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," *arXiv preprint arXiv:2406.01574*, 2024.

[41] M. S. A. Pal and M. Sankarasubbu, "Openbiollms: Advancing open-source large language models for healthcare and life sciences," 2024.

[42] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, and W. Xie, "Towards building multilingual language model for medicine," *Nature Communications*, vol. 15, no. 1, p. 8384, 2024.

[43] C. Christophe, P. K. Kanithi, P. Munjal, T. Raha, N. Hayat, R. Rajan, A. Al-Mahrooqi, A. Gupta, M. U. Salman, G. Gosal *et al.*, "Med42–evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches," *arXiv preprint arXiv:2404.14779*, 2024.

[44] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[45] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on Health, Inference, and Learning*. PMLR, 2022, pp. 248–260.

[46] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[47] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "Huatuogpt-o1, towards medical complex reasoning with llms," *arXiv preprint arXiv:2412.18925*, 2024.

[48] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[49] Q. Team, "Qwq: Reflect deeply on the boundaries of the unknown," November 2024. [Online]. Available: https://qwenlm.github.io/blog/qwq-32b-preview/

[50] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 2024.

[51] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[52] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "Gpqa: A graduate-level google-proof q&a benchmark," *arXiv preprint arXiv:2311.12022*, 2023.

[53] Q. Team, "Qwen2.5: A party of foundation models," September 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

Limitations and Future Work

Although AlphaMed achieves impressive results on multiple-choice QA tasks, its capabilities remain constrained by the closed-form nature of these benchmarks. Our evaluations are primarily conducted on existing mainstream medical QA datasets, all of which are close-ended and may not fully capture the spectrum of real-world clinical reasoning. Due to limitations in the current research landscape, it is challenging to systematically assess our model's performance on open-ended QA tasks, which not only lack well-established benchmarks but are also inherently subjective, often requiring human evaluation for meaningful assessment. In future work, we aim to design and release open-ended benchmarks that involve human-in-the-loop evaluation, enabling more comprehensive and nuanced assessments of reasoning and decision-making in medical LLMs.

Broader Impact

This work demonstrates that the reasoning capability of medical LLMs can be effectively incentivized using only multiple-choice QA data with minimalist rule-based RL, removing the need for SFT on costly distilled CoT data. By eliminating reliance on manual annotation and closed-source supervision, our approach substantially reduces the human effort and resources required for developing high-performing clinical models. However, the emerging reasoning processes in LLMs are inherently difficult to evaluate, as there is often no single "ground truth" reasoning path—especially in medicine, where multiple valid clinical justifications may exist for a single decision. Nonetheless, exposing these intermediate reasoning steps provides an important opportunity to observe and audit model behavior, ultimately encouraging the development of more transparent and trustworthy medical LLMs.

A Appendix

A.1 Difficulty Level Distribution

To explore how the difficulty level of training data affects model performance, we annotate each sample by its response consistency across five inference passes of `Llama3.1-8B-Instruct` [46]. Specifically, L1 denotes samples where the model answers all attempts correctly (easy), while L6 includes those where all predictions are incorrect (hard). Intermediate levels (L2–L5) indicate varying degrees of partial correctness. Tab. A.1 summarizes the distribution across MedQA[5], MedMCQA[6], and PubMedQA[7].

Table 2: Difficulty Level Distribution. L1 indicates samples where `Llama3.1-8B-Instruct` [46] predicts correctly in all 5 inference attempts (easiest), while L6 corresponds to samples where all predictions are incorrect (hardest). Intermediate levels (L2–L5) reflect partial correctness across attempts.

Dataset	Total	L1	L2	L3	L4	L5	L6
MedQA	10,178	1,970	1,471	934	697	713	4,393
MedMCQA	182,822	63,292	25,736	14,498	9,922	10,088	59,286
PubMedQA	211,268	97,790	41,604	18,596	10,759	9,217	33,303

A.2 Details of Evaluation Datasets

To thoroughly assess performance across varying levels of challenge, we evaluate on six medical QA benchmarks, grouped by challenge level:

Normal challenge level

[5]https://huggingface.co/datasets/GBaker/MedQA-USMLE-4-options-hf

[6]https://huggingface.co/datasets/openlifescienceai/medmcqa

[7]https://huggingface.co/datasets/qiaojin/PubMedQA

- MedQA [44]: A benchmark derived from US medical licensing exam questions, assessing clinical knowledge across a wide range of topics. Evaluation is based on the standard test split.
- MedMCQA [45]: A medical QA dataset based on entrance exams, designed to test foundational medical knowledge through multiple-choice questions. The official test split is used.
- PubMedQA [35]: A biomedical question answering dataset where models choose from three fixed options, yes, no, or maybe, based on associated research abstracts, emphasizing factual understanding in biomedical literature. The official test split is used.

Hard challenge level

- MMLU-ProM [40]: MMLU-ProM is the medical category subset of a broad multitask benchmark, focusing on professional-level medicine and related domains. Evaluation is conducted using the standard split established in [47].
- GPQA-M [37]: It represents the biomedical subset of a graduate-level QA benchmark, featuring expert-curated questions intentionally designed to resist superficial retrieval and demand deep analytical reasoning. The evaluation follows the split from [47].

Hard+ challenge level

- MedXpert [38]: A challenging benchmark designed to assess expert-level medical knowledge, clinical understanding, and complex reasoning. It covers diverse specialties and body systems, incorporates board-style exam questions, and is curated through expert review to ensure high difficulty, accuracy, and relevance to real-world medical decision-making.

A.3 Effect of LLM Backbones

To assess the generality of our proposed training pipeline and data design, we further apply the same minimalist rule-based RL approach, originally used for `Llama3.1-8B-Instruct`, to `Qwen2.5-7B-Instruct` [53]. After training, the resulting AlphaMed(7B) model achieves consistent improvements across all six benchmarks, as shown in Fig. 9. Notably, the gains are particularly substantial on the more challenging datasets, MMLU-ProM [36], GPQA-M [37], and MedXpert [38], demonstrating the robustness of our training strategy in enhancing medical reasoning. These results demonstrate that minimalist rule-based RL can incentivize reasoning capabilities and boost performance, exhibiting robustness across different backbone models.

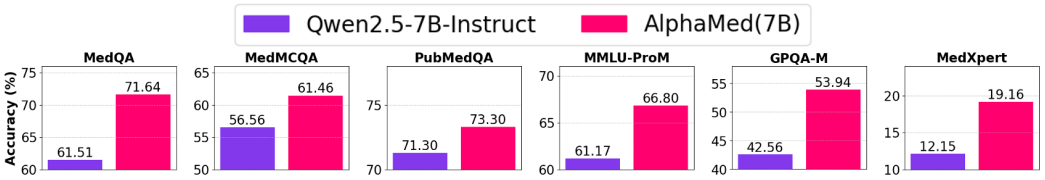

Figure 9: Performance comparison across six medical QA benchmarks. AlphaMed(7B) is initialized from `Qwen2.5-7B-Instruct` [53] and trained using our constructed training set and minimalist rule-based RL pipeline. It achieves consistent improvements over the base model on all benchmarks.

A.4 Success on Small LLM

To further evaluate the effectiveness of our minimalist RL pipeline, we apply it to a small language model, `Qwen2.5-3B-Instruct` [53]. As shown in Fig. 10, our approach consistently improves performance across all six medical benchmarks, including substantial gains on **MedQA (+11.55%)**, **GPQA-M (+19.19%)**, and **MedXpert (+4.10%)**. These results demonstrate that our RL framework can effectively incentivize reasoning capabilities even in smaller-scale models, and is not limited to large foundation models.

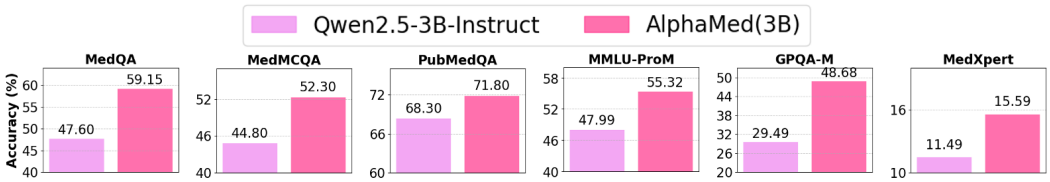

Figure 10: Performance comparison across six medical QA benchmarks. AlphaMed(3B) is initialized from `Qwen2.5-3B-Instruct` [53] and trained with our constructed dataset using a minimalist rule-based RL pipeline. It achieves consistent gains over the base model.

A.5 Qualitative Results

We present three examples predicted by our model trained with minimalist RL, demonstrating interpretable step by step clinical reasoning across diverse case types. In Fig. 11, the model correctly identifies inappropriate and potentially harmful options (e.g., use of NOACs in patients with mechanical heart valves) and adheres to guidelines by recommending bridging strategies based on patient risk factors and procedural context. In Fig. 12, it performs multi step numerical reasoning to derive absolute risk reduction (ARR) and relative risk (RR), showcasing its ability to integrate clinical knowledge with quantitative interpretation. In Fig. 13, the model applies structured reasoning to diagnose croup in a pediatric patient, identifying clinical features, linking them to pathophysiology, and reviewing radiographic findings, despite being supervised only on the final answer choice. This highlights the model's capacity for guideline aligned reasoning and emergent interpretability, even without supervision on intermediate reasoning traces.

Question:
A 78-year-old male patient visits the physician for a preoperative evaluation five days before scheduled abdominal hernia repair. Two years ago, he underwent mechanical mitral valve replacement for chronic primary mitral regurgitation. He has a history of hypertension and has been smoking one pack of cigarettes daily for 30 years. His current medications include warfarin, hydrochlorothiazide, and aspirin. Vital signs show a temperature of 37.1°C (99.3°F), pulse of 88/min, and blood pressure of 126/74 mm Hg. Physical examination reveals clear lung sounds, a mitral click on cardiac auscultation, and a soft, nontender abdomen. No other abnormalities are noted. Laboratory results are as follows:
Hemoglobin	13.1 g/dL
Leukocyte count	8900/mm³
Platelet count	230,000/mm³
Prothrombin time	23 seconds (INR = 2.1)
Creatinine clearance	102 mL/min
Urinalysis is within normal limits. A chest x-ray shows no abnormalities, and an ECG reveals mild left atrial enlargement. Regarding his medication management, which of the following instructions is most appropriate at this time?
Answer Choices:
A: Stop warfarin and start unfractionated heparin now, discontinue 4–6 hours prior to surgery
B: Stop warfarin now and start fondaparinux bridge therapy
C: Continue warfarin at half dose until surgery
D: Stop warfarin now and continue post-surgery
E: Stop warfarin now, start dalteparin therapy in 2 days, and discontinue 24 hours prior to surgery **(Ground Truth)**
F: Stop warfarin 3 days before surgery without bridging anticoagulation
G: Switch to rivaroxaban and discontinue 24 hours before surgery
H: Switch warfarin to dabigatran and discontinue 48–72 hours prior to surgery
I: Continue warfarin until 24 hours before surgery, then start enoxaparin bridge
J: Stop warfarin and aspirin now, resume both 48 hours post-surgery

Answer: **Step 1: Assess the patient's risk factors for thromboembolism and bleeding**
The patient has a mechanical mitral valve replacement, which is a high-risk condition for thromboembolism. However, the patient is also undergoing a low-risk surgery (abdominal hernia repair) and has been on warfarin with an INR of 2.1, which is within the therapeutic range.
Step 2: Determine the need for bridging anticoagulation
Given the high-risk condition of the mechanical mitral valve, bridging anticoagulation is typically recommended to prevent thromboembolic events during the period when warfarin is stopped. However, the risk of bleeding also needs to be considered, especially in surgical patients.
Step 3: Evaluate the options for bridging anticoagulation
Options include stopping warfarin and starting unfractionated heparin (UFH) or low molecular weight heparin (LMWH) like dalteparin or enoxaparin, or using fondaparinux.
The timing of stopping warfarin and starting the bridge therapy, as well as when to discontinue it before surgery, is crucial.
Step 4: Consider the specific instructions for each option
Option A suggests stopping warfarin and starting UFH, which is typically used in higher-risk situations or when rapid anticoagulation reversal is needed.
Option B suggests fondaparinux bridge therapy, which is an alternative to LMWH but has its own set of considerations.
Option C reduces warfarin dosage, which does not reliably prevent bleeding and may leave the patient inadequately anticoagulated. This approach is not supported by guidelines.
Option D omits bridging in a high-risk patient with a mechanical mitral valve, creating an unacceptably high risk of thromboembolism. This option is unsafe.
Option E suggests stopping warfarin, starting dalteparin in 2 days, and discontinuing 24 hours prior to surgery, which is a common approach for bridging in high-risk patients undergoing surgery.
Option F suggests stopping warfarin 3 days before surgery without bridging, which might be considered for lower-risk situations or surgeries with lower bleeding risk.
Option G uses rivaroxaban, a NOAC, which is contraindicated in patients with mechanical heart valves due to increased risk of stroke and valve thrombosis demonstrated in the RE-ALIGN trial.
Option H uses dabigatran, also a NOAC, which is contraindicated in mechanical valve patients for the same reasons as rivaroxaban. This option is unsafe and not guideline-supported.
Option I delays bridging too long and leaves a window of subtherapeutic anticoagulation. Bridging should be initiated earlier in high-risk patients.
Option J discontinues all antithrombotic therapy without bridging, creating an extended anticoagulation-free period in a high-risk patient, increasing the risk of thromboembolic events.
Step 5: Choose the most appropriate option based on the patient's profile and surgical risk
Given the patient's high-risk condition (mechanical mitral valve) and the relatively low-risk nature of the surgery, bridging anticoagulation is likely necessary.
The most common and recommended approach for someone with a mechanical heart valve undergoing surgery would be to stop warfarin and bridge with LMWH, resuming anticoagulation post-operatively when safe.
The final answer is: $\\boxed{E}$

Figure 11: Question and answer pair for Case 1. **Cyan text** highlights the final predicted choices. Green highlight are used to emphasize reasoning steps and key clinically key information.

Question: A clinical trial was conducted with 15,225 hypertensive patients allocated to the intervention group (new drug) and the control group (old drug), respectively. The results of the study are summarized in the following table. Calculate the absolute risk reduction (ARR) and relative risk (RR).

Group	Developed HT complications	Did not develop HT complications	Total
Control group (old drug)	1800	13,425	15,225
Intervention group (new drug)	1620	13,605	15,225

Answer Choices:
A: ARR = 10% and RR = 0.9
B: ARR = 1% and IZR = 9
C: ARR = 1% and RR = 0.9 **(Ground Truth)**
D: ARIZ = 10% and RR = 9

Answer: To calculate the absolute risk reduction (ARR) and relative risk (RR), we first need to determine the risk of developing hypertensive (HT) complications in both the control group (old drug) and the intervention group (new drug).

Step 1: Calculate the risk of developing HT complications in the control group and the intervention group.

•Risk in the control group = (Number of patients who developed HT complications in the control group) / (Total number of patients in the control group)

•Risk in the intervention group = (Number of patients who developed HT complications in the intervention group) / (Total number of patients in the intervention group)

Given data: Control group developed HT complications = 1800. Intervention group developed HT complications = 1620. Total subjects in each group = 15225. Risk in the control group = 1800 / 15225. Risk in the intervention group = 1620 / 15225

Step 2: Calculate the actual risks.

Risk in the control group = 1800 / 15225 ≈ 0.1182 or 11.82%

Risk in the intervention group = 1620 / 15225 ≈ 0.1064 or 10.64%

Step 3: Calculate the Absolute Risk Reduction (ARR).

ARR = Risk in the control group - Risk in the intervention group

ARR = 0.1182 - 0.1064 = 0.0118 or 1.18%

Step 4: Calculate the Relative Risk (RR).

RR = Risk in the intervention group / Risk in the control group

RR = 0.1064 / 0.1182 ≈ 0.9

Step 5: Match the calculated values with the given options.

•ARR ≈ 1.18% (approximately 1% when rounded)

•RR ≈ 0.9

The final answer is: $\boxed{C}$

Figure 12: Question and answer pair for Case 2. **Cyan text** highlights the final predicted choices. Green highlight are used to emphasize reasoning steps and key clinically key information.

Question: A 3-year-old boy presents to the emergency department with a 'cough-like-a-seal bark' and a high-pitched inspiratory noise that is audible without a stethoscope. His mother reports that his cough has worsened over the last few hours. The patient's blood pressure is 118/78 mm Hg, pulse is 90/min, respiratory rate is 35/min, and temperature is 38.3°C (101.1°F). On physical examination, the boy is sitting and leaning forward in apparent respiratory distress with suprasternal and intercostal retractions. Auscultation reveals inspiratory stridor without wheezing. He has a frequent barking cough and a hoarse voice when he speaks. What is a chest X-ray likely to show?

Answer Choices:

A: Increased interstitial markings
B: Lobar consolidation in the lingual
C: Thumbprint sign on the lateral image
D: Steeple sign **(Ground Truth)**

Answer: To answer this question, let's reason through the case step by step.
Step 1: Identify the clinical presentation.
A 3-year-old boy presents with a "cough-like-a-seal bark" and a high-pitched inspiratory noise. The cough has worsened over the last few hours. Physical examination reveals that he is sitting and leaning forward in apparent respiratory distress with suprasternal and intercostal retractions. Inspiratory stridor is present without wheezing. He also has a frequent barking cough and a hoarse voice.
Step 2: Diagnose the condition.
The combination of a barking cough, hoarseness, and inspiratory stridor in a child is characteristic of croup, which is typically caused by viral infections such as parainfluenza virus.
Step 3: Understand the radiographic findings in croup. Croup is characterized by inflammation of the larynx, trachea, and bronchi, particularly the subglottic area. This inflammation leads to edema and narrowing of the subglottic airway. On a chest X-ray, croup typically shows a "steeple sign."
Step 4: Review the chest X-ray findings for each option.
A: Increased interstitial markings – This is more characteristic of interstitial lung disease or pneumonia, not croup.
B: Lobar consolidation in the lingula – This is more characteristic of pneumonia, particularly in the lower lobe, not croup.
C: Thumbprint sign on the lateral image – This is a rare finding in croup and is not as specific as the steeple sign. It is more commonly associated with other conditions like tracheal stenosis.
D: Steeple sign – This is the characteristic finding in croup. The "steeple sign" refers to the appearance of the narrowed subglottic airway on a lateral view of the neck, which resembles the top of a steeple.
The final answer is: $\boxed{D}$

Figure 13: Question and answer pair for Case 3. **Cyan text** highlights the final predicted choices. Green highlight are used to emphasize reasoning steps and key clinically key information.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction accurately represent the contributions and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Refer to the start of Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work mainly includes empirical contributions.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental configurations in Sections 4 and 5, and Appendix. Our code and pretrained weight will be released after acceptance.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: Our experiments are all conducted on publicly accessible datasets and models, and all experiment details are illustrated in Sections 4.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are illustrated in Sections 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following prior works, we report performance generated with temperature 0 and greedy sampling during LLM inference to ensure deterministic outputs across runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to the first part of Sections 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: Replace by [Yes]

   Justification: This work is conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the second section of Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: This work uses only verified datasets and does not directly involve patient-related scenarios.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use all public datasets as mentioned in Section 5.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: There is no new assets released in this work.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This work has no human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This work has no human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We perform reinforcement learning-based post-training to directly optimize the reasoning capabilities of LLMs without relying on supervised fine-tuning or chain-of-thought annotations.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.