Med-U1: A Unified Reinforcement Learning Framework for Controllable Medical Reasoning in LLMs

Anonymous ACL submission

Abstract

Medical Question-Answering (QA) encompasses a broad spectrum of tasks, including multiple choice questions (MCQ), open-ended text generation, and complex computational reasoning. Despite this variety, a unified framework for delivering high-quality medical QA has yet to emerge. Although recent progress in reasoning-augmented large language models (LLMs) has shown promise, their ability to achieve comprehensive medical understanding is still largely unexplored. In this paper, we present Med-U1, a unified framework for robust reasoning across medical QA tasks with diverse output formats, ranging from MCQs to open generation and computation tasks. Med-U1 employs pure large-scale reinforcement learning with mixed rule-based binary reward functions, incorporating a length penalty to manage output verbosity. With multi-objective reward optimization, Med-U1 directs LLMs to produce controllable and verifiable reasoning chains. Empirical results reveal that Med-U1 significantly improves performance across multiple challenging Med-QA benchmarks, surpassing even larger specialized and proprietary models. Furthermore, Med-U1 demonstrates robust generalization to out-of-distribution (OOD) tasks. Extensive analysis presents insights into training strategies, reasoning chain length control, and reward design for unified reasoning-augmented medical LLMs.

1 Introduction

Recent advances in reinforcement learning (RL) based long-form reasoning for medical scenarios have significantly improved model performance of large language models (LLMs) on complex tasks (Zhang et al., 2024; Griot et al., 2025; Liévin et al., 2024; Liu et al., 2024a; Jiang et al., 2025; Wang et al., 2025). Unlike domains such as mathematical reasoning or code generation, where answers are typically verifiable and label formats are relatively consistent (Jaech et al., 2024; Team et al., 2025; Team, 2025), medical reasoning spans a wide range of types, including conditional inference, numerical computation, open-ended generation, and etc., each exhibiting substantial variation in difficulties and output formats (Khandekar et al., 2024; Zuo et al., 2025; Chen et al., 2024; Liu et al., 2024b). Beyond standard multiple choice questions (MCQs) with verifiable answers, medical tasks also include computational reasoning for patient care (e.g., hospital stay duration or fluid maintenance) and open-ended text generation with short phrases or long-form text, requiring deep understanding and varied lengths of reasoning, as illustrated in Figures 4.

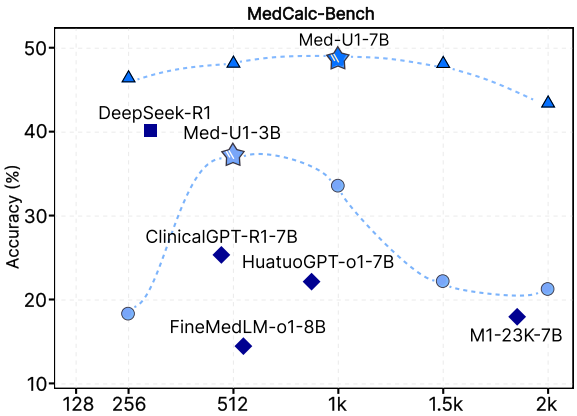

Figure 1: Med-U1 is a length-controllable reasoning model trained with a length-aware reward, achieving superior performance compared to other models. It allows users to specify the desired reasoning length based on task requirements.

Recent research highlights that large-scale RL, such as DeepSeek-R1-Zero (Guo et al., 2025), can achieve promising performance using binary rule-based reward signals without requiring extensive human-annotated data for Supervised Fine-Tuning (SFT) or Chain-of-Thought (CoT). Several efforts have adapted RL-style training paradigms to the medical domain. (Chen et al., 2024) utilized reasoning traces generated from MCQ datasets to per-

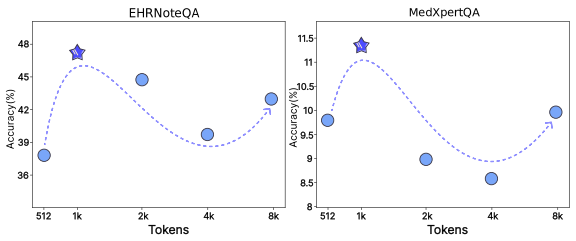

Figure 2: Accuracy under different inference lengths.

form SFT, followed by proximal policy optimization (PPO) for RL. Similarly, (Yu et al., 2025) employed multi-stage SFT with synthetic data and long-form reasoning samples, enhancing reasoning capabilities through Direct Preference Optimization (DPO). However, these methods still rely heavily on large-scale supervised data and underexplore pure RL's potential to improve reasoning. More critically, they focus on tasks with verifiable answers, such as MCQs, limiting their applicability to diverse medical reasoning tasks and real-world clinical scenarios (e.g., computation or generation).

Moreover, efficient inference with high accuracy is critical across diverse computational environments, particularly given the varying difficulty and complexity of medical tasks (Huang et al., 2025c). Recent studies indicate user preferences for controllable reasoning lengths, favoring detailed explanations for complex queries and concise responses for simpler ones (Aggarwal and Welleck, 2025). However, existing methods rarely address this need, especially in the medical domain. For instance, (Huang et al., 2025b) introduced a "budget forcing" technique to compel continued reasoning. While effective, this approach incurs significant computational overhead. Results in Figure 1 suggest that flexible, dynamic control of reasoning length could reduce inference costs and enhance performance, with Med-U1-3B achieving almost optimal results at 512 tokens.

In this paper, we propose **Med-U1**, a unified and pure RL framework with controllable-length reasoning across diverse medical QA tasks, ranging from MCQs to complex generation and computation tasks. We propose a mixed medical task-specific reward function, which extends the binary rule-based rewards to effectively guide training in a variety of medical contexts. Specifically, we explore reward optimization for different medical reasoning tasks, including MCQs, numerical value computation, and open-ended text generation. Furthermore, we design a reward to incorporate length constraints to dynamically regulate output verbosity. This multi-objective reward design enables LLMs to improve test-time performance via verifiable reasoning while maintaining controllability of reasoning length. Experimental results on 5 challenging benchamarks demonstrate that Med-U1 substantially improves model performance across diverse tasks. The Med-U1-7B surpasses larger-scale models such as Qwen-32B, as well as strong proprietary baselines. Furthermore, Med-U1 exhibits strong out-of-distribution (OOD) generalization capabilities, as evidenced by its superior performance on the MMLU-Pro-health benchmark. We further conduct ablation studies on different initialization strategies and reward guidance configurations, revealing that medical tasks exhibit heightened sensitivity to variations in RL paradigms. In addition, we explore the task-dependent optimal inference lengths across diverse scenarios and analyze the effects of different reward signal designs. Our contributions are:

- We present a unified medical framework driven by verifiable reward signals, tailored for diverse medical tasks, achieving strong performance across both in-distribution and OOD tasks, demonstrating its robustness and adaptability in diverse clinical settings.
- We introduce a length constraint reward to dynamically balance task complexity and reasoning cost, and analyze how different medical tasks exhibit varying preferences and sensitivities to the reasoning length.
- We demonstrate that Med-U1 achieves leading performance across a wide range of medical QA tasks, outperforming both specialized medical reasoning models and general-purpose models, even under limited reasoning budgets.

2 Preliminaries and Related Work

2.1 Medical Reasoning of LLMs.

Previous medical LLMs have relied mainly on two approaches: task-specific prompting and further fine-tuning with medical data. However, the dependence on large-scale annotated datasets and the sensitiveness to task formulation significantly limit their generalizability across diverse medical subtasks (Jiang et al., 2024). Although the o1 paradigm has demonstrated remarkable success in

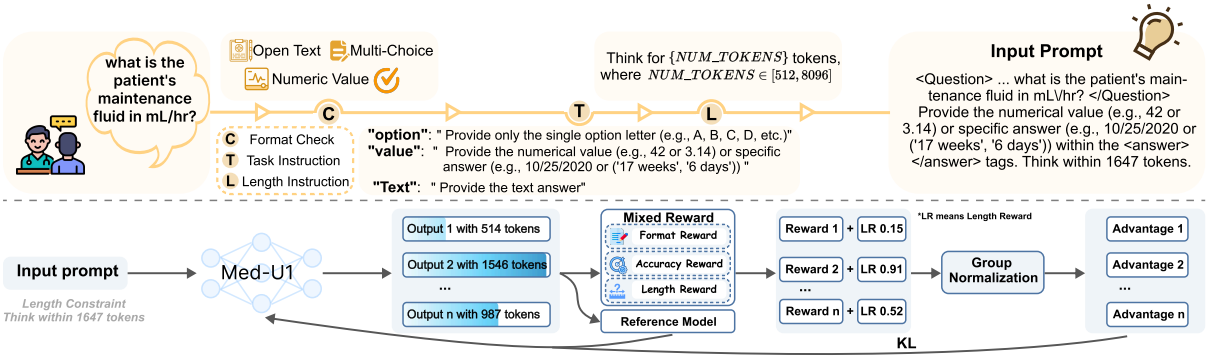

Figure 3: Overview of the Med-U1 training framework. **Upper:** The generation of input prompt based on different tasks formats. **Lower:** The policy is optimized using a outcome-driven Mixed Reward strategy, which encourages both accurate predictions and length-matched outputs.

general domains, few studies have systematically investigated its application to deliberative chain-of-thought reasoning in clinical settings (Yu et al., 2025). HuatuoGPT-o1 (Chen et al., 2024) adopted a fine-tuning strategy that constructs complex reasoning trajectories, further enhanced through reinforcement learning. FineMedLM-o1 (Yu et al., 2025) enhanced reasoning capabilities through multi-stage supervised fine-tuning and direct preference optimization, and further incorporates Test-Time Training (TTT) to promote domain adaptation and improve the reliability of reasoning. Med-RLVR (Zhang et al., 2025) explored the feasibility of Reinforcement Learning from Verifiable Rewards (RLVR) in the medical domain, highlighting the potential of pure reinforcement learning to improve model reliability and reasoning accuracy.

However, existing approaches predominantly focus on single tasks, with the reward constructed in a multiple choice format. As a result, they lack a systematic exploration of generalizing across heterogeneous medical tasks and supervision formats. Meanwhile, users face varying reasoning budget demands across tasks, such as the need to control inference length during test-time. Despite this, unified training paradigms across heterogeneous medical tasks have yet to be adequately explored in real-world clinical settings.

2.2 Length Control in LLMs.

Controlling the output length of LLMs is a critical practical consideration across various generation tasks (Kang et al., 2025; Xia et al., 2025; Luo et al., 2025). Prior efforts have explored different strategies to address this issue, Butcher et al. (Butcher et al., 2025) manipulated positional encodings to enforce precise sequence lengths; Yuan et al. (Yuan et al., 2024) trained models on instruction-style data explicitly annotated with target output lengths; and Singhal et al. (Singhal et al., 2023) modified training objectives to directly enforce length constraints. However, these methods have predominantly focused on general-purpose text generation or instruction-following contexts. Meanwhile, several works have exploited additional test-time inference to achieve performance gains. S1 (Muennighoff et al., 2025) introduced a strict token budget enforcement mechanism, referred to as "budget-forcing", which enables models to truncate outputs or insert a special token "Wait" to explicitly request continuation. Similarly, L1 (Aggarwal and Welleck, 2025) proposed rule-based reward functions that penalize outputs that violate length constraints, enabling models to be trained with sensitivity to predefined length requirements.

However, when dealing with complex medical tasks, a natural question arises: Is simple test-time scaling still effective? To investigate this, we conduct a preliminary evaluation using L1-Qwen-7B (Aggarwal and Welleck, 2025) on EHRNoteQA and MedXpertQA, as shown in Figure 2. The results indicate that accuracy on both datasets exhibits fluctuations rather than a linear improvement with increased reasoning length, suggesting potential differences in reasoning demands across distinct medical tasks.

3 Method

To bridge the aforementioned gap, we propose Med-U1, as presented in Figure 3. Based on output formats, we categorize existing medical QA tasks into three types: multi-choices, numeric value and

open text, each might contain further fine-grained formats. Given a generated instruction and question, Med-U1 is trained in a unified manner under a mixed reward scheme comprising format verification, correctness validation, and length constraint.

3.1 Task-specific Reward

The reward signal serves as the primary driver for model optimization in RL. Unlike mathematics and coding, where answers are typically verifiable, medical tasks exhibit diverse output formats. As shown in Figure 3, in this paper, we classify these formats into three categories: (1) multiple choice responses with explicit lettered options, (2) numerical values, such as exact numbers short phrases, and (3) open-ended free-text responses. To address this diversity, we propose a task-specific reward function with two components: a format reward that evaluates structural compliance of outputs and a correctness reward that assesses content accuracy.

Format Reward. Regardless of the task type, the model is required to include its reasoning process within `<think></think>` tags and its predicted answer within `<answer></answer>` tags in order to obtain a higher format reward. We employ regular expressions to extract the predicted answer from the output, which is then compared to the ground truth for further correctness evaluation. The format reward R_{format} is computed as:

$$R_{format} = \begin{cases} 1, & \text{if format is correct,} \\ -1, & \text{if format is incorrect.} \end{cases} \quad (1)$$

Correctness Reward. Only after the response format has been verified do we proceed to assess the quality of the model's prediction. The answer quality score $R_{correct}$ is computed as follows in MCQs and numeric computation tasks:

$$R_{correct} = \begin{cases} 1, & \text{if answer is correct,} \\ -1, & \text{if answer is incorrect,} \end{cases} \quad (2)$$

and for open-ended generation tasks, the correctness reward $R_{correct}$ is computed as $R_{correct} = Score(pred, ref)$, where the score is clipped to [-1, 1] due to the challenge of capturing fine-grained differences between discrete outputs.

We explore several task-specific metrics to verify answers according to the nature of the task:

- **MCQs (Options):** The evaluation is the same as math benchmarks. The predicted option is extracted from the `<answer></answer>` tags using regular expressions and directly compared to the ground truth (true o false).
- **Numeric Value Computation:** In ths category, some tasks involve numerical reasoning under specified constraints and output a number (e.g., 42 or 3.14). In other tasks, the prompt provides contextual reasoning cues, and the model is required to output short phrases in specific formats, such as treatment durations (e.g., "17 weeks", "6 days") or dates (e.g., "10/25/2020"). For such tasks with numeric value answers, $R_{correct}$ is determined by verifying that the predicted value lies within the allowable range (that is, lower limit $\leq$ predicted value $\leq$ upper limit), or by performing a direct character-level match against the ground truth.
- **Open-Ended Text Generation:** Open-ended QA represents the most common task in real-world clinical scenarios, for example, generating diagnostic conclusions or treatment plans based on the information provided. These tasks are particularly challenging due to the inherent variability in output formats. To assess the quality of generated content, metrics such as BLEU (Papineni et al., 2002) and Rouge-L (Lin, 2004) are commonly used.

Failure Case of BLEU Reward

Reference: At term, the ratio of the weight of the fetus to the weight of the placenta is typically about 6:1.
Model Output: 6:01. **Bleu:** 0.01.

However, we observe the above faluire case (in black box) of BLEU reward in our preliminary experiments. Specifically, model outputs within the `<answer></answer>` tags are often concise factual spans, resulting in token-level differences that BLEU fails to reward correctly due to its emphasis on partial n-gram overlap. In cases where the model generates a concise yet accurate description, BLEU may yield misleadingly low scores. To address this issue, we adopt Exact Match Score (EMS) as the main reward signal during training. EMS better reflects the correctness of short structured outputs by explicitly measuring word-level match. To provide a softer reward gradient and accommodate near-miss cases, we

further incorporate Rouge-L as an auxiliary reward. Rouge-L captures the longest common subsequence between candidate and reference, thus encouraging reward even when full exact matches are not present. In summary, we define a mix evaluation criterion that takes the average of the Rouge-L score and the EMS, assigning equal weight to each component. In addition, we explore training variants that either Rouge-L or EMS is used independently as the sole reward signal.

3.2 Target Length Matching Reward

Unconstrained reasoning models often suffer from overthinking, where excessive and unnecessary reasoning is performed for simple tasks, leading to computational inefficiency. Conversely, such models may also terminate reasoning prematurely in complex scenarios that require deeper inference. In medical domains with varying levels of task complexity, achieving fast and accurate reasoning with limited computational resources is critical. This requires that models' reasoning length can be dynamically aligned with user-specified requirements.

Given a question q and a target length l_{gold}, the model is encouraged to produce a response y with a total thinking length l_y such that the absolute difference $|l_y - l_{gold}|$ is minimized to maximize the reward R_{length}. We define the R_{length} as follows:

$$R_{length} = \max\left(-1, \min\left(1, 1 - \frac{|l_y - l_{\text{gold}}|}{\alpha}\right)\right), \quad (3)$$

where α is a normalization factor related to the max output length, and the length reward is ultimately clipped to the interval [-1,1]. By combining the length-based reward with the previously established reward, we encourage the model to generate correct answers through faithful reasoning while simultaneously adhering to user-defined length constraints. This joint reward formulation promotes both correct and efficient reasoning in accordance with user expectations.

3.3 Policy Optimization

As shown in Figure 3, the final mixed reward combines three terms: a format reward R_{format}, a correctness reward $R_{correct}$ and a target length matching reward R_{length}. R_{length} is optional; when omitted, the model is encouraged to reason freely based on its own understanding of the question. The reward function serves a dual purpose: (a) it encourages the model to follow the required format and produce correct answers while implicitly favoring concise reasoning trajectories when shorter outputs are requested by the user; and (b) it consistently incentivizes the model to match the specified target length, thereby facilitating the exploration of optimal reasoning lengths for different task types.

We use the Group Relative Policy Optimization (GRPO) algorithm to train the model with the final mixed reward. In each training step, for the given question q, a group of candidate outputs $O = \{o_1, o_2, \cdots, o_G\}$ are sampled from the policy model $\pi_{\theta_{old}}$. The advantage $A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$ is calculated using the mixed rewards $\{r_1, r_2, \cdots, r_G\}$. Then the following objective function is maximized to optimize π_θ:

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q),\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)} \\ \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \\ \text{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon,\, 1+\varepsilon \Big) A_i \Big) \\ - \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \Bigg], \end{aligned} \quad (4)$$

where ε and β are hyperparameters controlling the PPO clipping threshold and the weight of the Kullback–Leibler (KL) divergence penalty (Schulman et al., 2017; Shao et al., 2024), respectively. ε determines the permissible range for policy updates, while β controls the KL divergence penalty.

4 Experiments

4.1 Experimental Settings

Dataset and Benchmarks. We select medical datasets which are categorized into different task scenarios and cover a variety of label formats: EHRNoteQA (Kweon et al., 2024), MedXpertQA (Zuo et al., 2025), MEDiQ (Li et al., 2024), MedCalc-Bench (Khandekar et al., 2024), medical-o1-reasoning-SFT (HuaTuoGPT) (Chen et al., 2024). More details on the above benchmarks are provided in Appendix. It is worth noting that we excluded intuition-based datasets, as they typically provide limited contextual information in the prompts and are therefore less suitable for supporting long-form reasoning. For the OOD test, we use the health subset of MMLU-Pro (Wang et al., 2024) which consists of 818 challenging multiple choice questions related to healthcare and medicine.

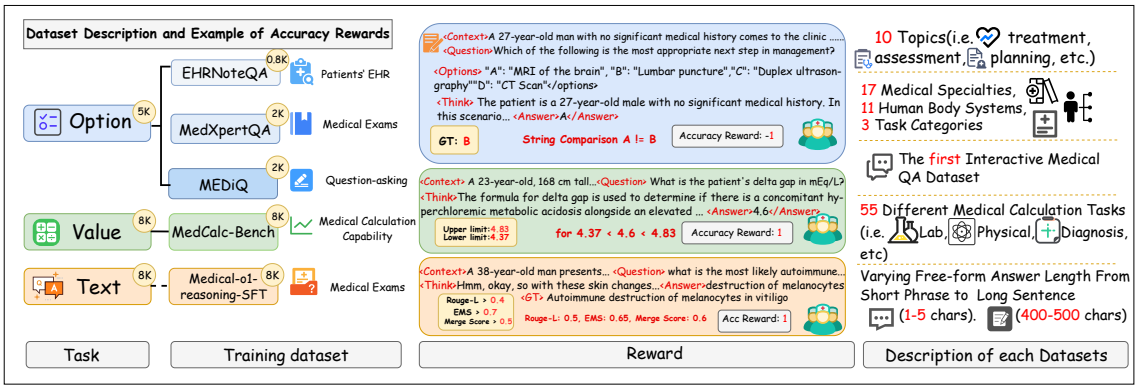

Figure 4: Datasets overview for training and in-distribution evaluation, with examples on reward computation.

Baselines. Our primary baselines encompass leading proprietary models, namely GPT-4.1, DeepSeek-R1, and Gemini 2.5 Flash, as well as advanced open-source models such as the Qwen2.5 series (Yang et al., 2024) and Llama-3.1-70B-Instruct. We also include strong domain-specific baselines, including HuatuoGPT-o1, FineMedLM-o1, ClinicalGPT-R1 and M1-23K, which are specifically tailored for medical applications. Proprietary models are accessed via their APIs. More evaluation details can be found in Appendix.

Evaluation Metrics. For tasks with available ground truth labels, we employ accuracy as the evaluation metric. In contrast, for open-ended text generation tasks, we evaluate the output quality using a set of complementary metrics: Exact Match Score (EMS) and Rouge-L, which together provide a comprehensive assessment of lexical overlap, and sequence-level correspondence.

Training Details. Our implementation is based on the verl framework. We utilize Qwen2.5-3B/7B-Instruct as the backbone model for training. During training, we configure a batch size of 4 and utilize 8 rollouts per prompt within the GRPO algorithm. We employ a constant learning rate of 5e-7 and set the sampling temperature to 1.0. We set the KL penalty coefficient to 0.001. All models are trained for 1 epoch on 4 NVIDIA A100 80G GPUs.

4.2 Main Results

In-Distribution Performance. Results demonstrate that Med-U1 achieves substantial improvements over its corresponding base versions. Furthermore, it consistently outperforms existing medical reasoning models of comparable scale. It exhibits competitive performance even when compared to significantly larger open-source and proprietary models (Table 1). On relatively simple benchmarks such as MEDiQ and EHRNoteQA, most baseline models achieve over 40% accuracy. However, on the more challenging MedXpertQA benchmark, our Med-U1-7B model attains an accuracy of 17.74%, outperforming nearly all open-source baselines and rivaling 70B-scale general-purpose model and proprietary models. On MedCalc-Bench, which requires higher precision and numerical sensitivity in arithmetic reasoning, the pure RL-based training paradigm significantly enhances the model's inferential and computational abilities, with Med-U1-7B reaching 57.55% accuracy. In the open-ended generation task, Med-U1 outperforms other methods in both Rouge-L and EMS metrics. We attribute its superior EMS performance primarily to the format reward introduced during training, which encourages the model to place the predicted answer precisely within the `<answer></answer>` tags, facilitating better alignment with the ground truth.

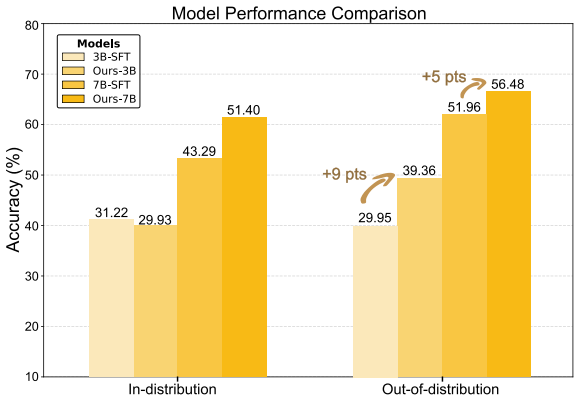

Figure 5: Comparing the proposed method and SFT on in-distribution and out-of-distribution tasks.

Out-of-Distribution Performance. As shown in

Baselines	MEDiQ	EHRNoteQA	MedXpertQA	AVG_{option}	MedCalc-Bench	medical-o1-reasoning-SFT	
						Rouge-L	EMS
Leading Proprietary Models							
GPT-4.1-nano	64.84	56.16	11.02	34.60	27.53	20.78	64.41
Gemini-2.5-flash-lite	85.94	56.32	31.02	51.16	38.87	24.24	76.48
DeepSeek-R1	78.91	63.22	20.00	44.64	40.16	27.45	41.52
Medicine-Specific Reasoning LLMs							
Qwen2.5-3B-SFT	41.41	41.38	13.06	26.33	33.90	32.42	45.55
Qwen2.5-7B-SFT	64.84	42.53	12.65	32.89	43.34	25.32	63.44
HuatuoGPT-o1(7B)	72.66	44.83	14.29	36.39	22.11	14.97	21.92
FineMedLM-o1(8B)	60.16	40.23	10.61	30.07	14.31	20.00	35.58
ClinicalGPT-R1(7B)	56.25	43.68	11.84	30.28	18.59	22.78	32.79
M1(7B)	57.03	47.13	13.06	31.79	18.69	25.54	39.13
General Purpose LLMs							
Qwen2.5-3B-Instruct	53.12	37.93	9.39	27.01	15.41	17.22	52.87
Qwen2.5-7B-Instruct	60.94	39.08	8.16	28.76	22.66	19.98	60.56
Qwen2.5-32B-Instruct	70.31	44.83	8.57	32.69	33.30	29.83	52.07
Qwen2.5-72B-Instruct	53.91	29.89	6.12	23.98	33.67	25.80	69.38
Llama-3.1-70B-Instruct	78.91	57.47	18.37	42.69	15.85	21.17	57.63
The proposed							
Med-U1-3B (Ours)	50.78	55.17	12.90	31.48	39.07	39.69	69.80
Med-U1-7B (Ours)	65.63	52.87	17.74	37.77	57.55	39.19	73.06

Table 1: Performance comparison on in-distribution medical tasks. We report average accuracy (AVG_{option}) across 3 MCQ tasks. Open-source baselines are categorized by accessibility and domain specificity. The results are reported in terms of accuracy (%) and percentage-based scores.

	Nutrition	C-M	C-K	P-M	H-A	Anatomy	M-G	Virology
Qwen2.5-3B-SFT	26.26	18.75	27.78	22.05	20.93	16.46	27.78	43.48
Med-U1-3B	43.58	47.92	45.83	37.01	33.72	24.05	53.70	36.96
Qwen2.5-7B-SFT	52.17	41.67	47.22	42.91	30.23	34.18	46.32	**52.17**
Med-U1-7B	**60.34**	**64.58**	**61.11**	**58.66**	**40.70**	**48.10**	**61.11**	**52.17**

Table 2: Performance comparisons across different medical subdomains, where C-M denotes College Medicine, C-K denotes Clinical Knowledge, P-M denotes Professional Medicine, H-A denotes Human Aging, and M-G denotes Medical Genetics. Results are reported in accuracy (%).

Table 2, we evaluate eight OOD medical scenarios, with the overall results presented in Figure 5. RL configurations (Med-U1) outperform the SFT baseline, particularly in OOD settings, with the improvements on 3B model can reach up to 9 percentage points. The finding suggests that RL improves generalization, which has been observed in the general domain, also holds in medical tasks. While the use of `<think></think>` tags may lead the model to reason in reliable patterns, we argue that the primary performance gains stem from the RL paradigm itself, especially given the inherent randomness in OOD data distributions. Several scenarios such as College Medicine and Medical Genetics, where the input often contains richer context to support reasoning, Med-U1-7B shows a substantial performance gain of approximately 19 percentage points compared to the SFT baseline. This aligns with the intuition behind online reinforcement learning: by evaluating and iteratively improving upon self-generated outputs, the model primarily learns how to solve rather than merely how to fit annotated data distribution.

5 Findings and Analysis

Appropriate Reasoning Length. While test-time scaling methods have been shown to improve model performance in general-domain tasks such as mathematics and code generation, we observe that this trend does not consistently hold in the medical domain. Specifically, as shown in Table 3, increasing the reasoning length does not lead to significant performance gains. Most tasks reach a local optimum around a reasoning length of 512 tokens, and the overall performance curve resembles a bell-shaped (i.e., approximately normal) distri-

	MEDiQ	EHRNoteQA	MedXpertQA	MedCalc
256	**52.34**	49.42	11.43	18.39
512	50.00	**52.87**	**14.28**	**37.97**
1024 (1k)	47.65	50.57	11.02	33.59
1.5k	47.65	42.53	11.02	22.66
2k	45.31	45.98	9.79	21.17
3k	43.75	47.13	9.79	22.86
4k	47.65	40.23	9.39	21.57
8k	48.44	47.13	10.20	20.18

Table 3: Accuracy of the Med-U1-3B output on different datasets when constrained to specific token lengths during inference. Results are reported in accuracy (%).

Reward		Score	
Rouge-L	EMS	Rouge-L	Exact Match
✔		**40.32**	61.46
	✔	24.06	61.75
✔	✔	39.69	**69.80**

Table 4: Models trained with different reward formulations on open-ended generation tasks.

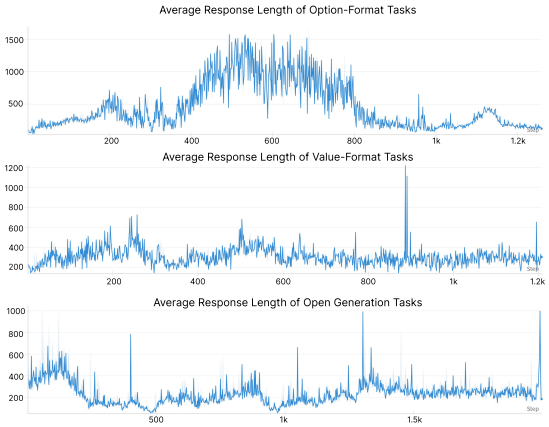

Figure 6: Average response length across various tasks over training steps. While the absolute lengths vary between tasks, the trends demonstrate a broadly consistent pattern of evolution during training.

bution. Specifically, for simpler benchmarks such as MEDiQ, extending the reasoning length actually leads to a decrease in prediction accuracy. We hypothesize that these tasks rely more on intuitive prediction than on step-by-step reasoning, and that excessive reasoning may amplify consistency errors and hallucinations in model outputs.

Impact of Reward Metric Selection. As described in Section 3.1, we explore three reward formulations for open-ended generation tasks: using only Rouge-L, using only EMS, and a Mix metric. As shown in Table 4, the choice of reward function has a substantial impact on the quality of the final model outputs, highlighting the sensitivity of RL training to the reward design. the format reward encourages the model to generate concise and precise response within `<answer></answer>`, making EMS particularly well suited for tasks requiring structured outputs. However, relying solely on EMS as the reward signal tends to overfit the model to a small subset of short target labels, thereby impairing generalization performance on the test set. Thus, the Mix metric balances Rouge-L and EMS with equal weights, preserving competitive performance compared to using only Rouge-L, and achieving superior EMS scores, demonstrating improved output quality and better generalization.

Emergence and Evolution of Thinking Patterns in Medical Tasks. By observing the training dynamics, we provide several insights into model adaptation and the emergence of reasoning. The response length initially increases slightly and then decreases, reflecting the transition of the model response from balancing R_{format} and $R_{correct}$ to adopting a more faithful `<think><answer>` reasoning pattern. A subsequent rise and fall in length indicates a shift from exploratory, complex reasoning toward more stable and direct analytical prediction. Variations in the amplitude of response length fluctuations across tasks suggest differences in the distribution of effective reasoning signals within the training data for each task.

Case Study of Different Training Paradigm. To better understand how output behaviors vary across training paradigms (Base, SFT, RL) and model scales, we conduct a qualitative case study. As shown in the Appendix, after obtaining patient background and the clinical query, the model must distinguish between relevant information (e.g., body weight) and irrelevant distractors (e.g., blood pressure, pulse). While all models begin with generally correct reasoning, some tend to second-guess themselves, leading to errors—likely due to suboptimal sampling and limited domain knowledge. In contrast, SFT and RL-trained models maintain more consistent and reliable reasoning.

6 Conclusion

In this work, we introduce Med-U1, a unified and RL-based framework tailored for robust reasoning across diverse medical question-answering tasks. Empirical evaluations across multiple Med-QA benchmarks demonstrate that Med-U1 surpasses existing state-of-the-art models, and exhibits strong generalization capabilities to OOD tasks.

Limitations

While this study provides promising evidence that reasoning length plays an important role in medical tasks, the current exploration is limited to a practical range constrained by available computational resources. Future work could further investigate longer reasoning budgets to verify whether additional gains may emerge beyond the explored range. In addition, open-ended medical reasoning involves more diverse and complex linguistic structures. Although the present reward design demonstrates effectiveness across multiple task types, extending it toward more fine-grained, clinically grounded objectives represents a valuable direction for future research.

Ethical Considerations

Med-U1 aims to provide a unified reinforcement learning framework for advancing medical reasoning in large language models. To date, no identifiable ethical or societal risks associated with Med-U1 have been observed. All experiments were conducted exclusively on publicly available medical question-answering benchmarks, and all base models used in this work are open-source and accessible through Hugging Face or publicly documented APIs. No private or patient-identifiable data were used at any stage of this research.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.

Bradley Butcher, Michael O'Keefe, and James Titchener. 2025. Precise length control for large language models. *Natural Language Processing Journal*, page 100143.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025a. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025b. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Sibo Song, Yang Feng, Jimeng Sun, Jian Wu, and Zuozhu Liu. 2025. Capo: Reinforcing consistent reasoning in medical decision-making. *Preprint*, arXiv:2506.12849.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *Preprint*, arXiv:2404.10237.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24312–24320.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, and 1 others. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024a. MedCoT: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024b. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning, march 2025. *URL https://qwenlm. github. io/blog/qwq-32b*, page 9.

Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. 2025. V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 658–668, Cham. Springer Nature Switzerland.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. 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*.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2024. Following length constraints in instructions. *arXiv preprint arXiv:2406.17744*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. Medrlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Details of Experiments

A.1 Datasets and Benchmarks

To ensure comprehensive and reliable training and evaluation, we select a suite of recently introduced datasets that are well-suited for assessing long-form reasoning, with all sources explicitly documented. EHRNoteQA is constructed from the MIMIC-IV electronic health record (EHR) database and contains 962 question–answer pairs, each associated with a distinct patient's discharge summary. EHRNoteQA includes questions that require information across multiple discharge summaries and covers ten diverse topics. MedXpertQA serves as a challenging and comprehensive benchmark designed to evaluate expert-level medical knowledge and advanced clinical reasoning skills. MEDiQ is originally developed to assess the question-asking capabilities of LLMs. However, due to the presence of rich contextual information that supports complex reasoning, we include it for the evaluation of reasoning abilities. MedCalc-Bench comprises more than 1,000 manually verified instances drawn from 55 distinct medical calculation tasks. It simulates real-world situations where physicians require quantitative reasoning and rule-based computations for evidence-based decision support. Medical-o1-reasoning-SFT is a synthetic dataset constructed using GPT-4o, where solutions to verifiable medical problems are automatically generated and validated by a medical verifier, enabling evaluation of multi-step reasoning and factual correctness. In the main paper, we summarize the number of samples used from each dataset during training, along with the subset of cases involved in reward computation. The data sources are also annotated to ensure that all selected datasets are credible and clinically valuable. Detailed Data Examples can be seen in Figure 10.

A.2 Description of medical-o1-reasoning SFT dataset.

In our text-format task, we have carefully considered the diversity of medical scenarios and the richness of medical answer formats. To achieve this, we utilized the medical-o1-reasoning-SFT dataset, which integrates the original MedQA-USMLE and MedMCQA datasets and transforms them into free-form answer formats. This transformation not only increases the difficulty of answering questions but also better reflects real-world medical problem-solving scenarios.

The medical-o1-reasoning-SFT dataset encompasses over 20 distinct medical disciplines, including Biochemistry, Anatomy, Physiology, and more. To further enhance the complexity of our task, we selected 10K challenging questions from this dataset. The answer length distribution of these questions, as shown in Figure 7, is diverse, ranging from short phrases (under 5 characters) to long sentences (over 400 characters). Such a wide range of answer formats significantly increases the difficulty of our text-format task and highlights the advantages of using RL, which can better handle the complexity and variability of medical questions and answers, enabling the model to provide more accurate and diverse responses in different medical scenarios.

A.3 Evaluation Details

When evaluating model performance on the test set, we deployed open-source models locally using frameworks such as vLLM or HuggingFace implementations. We use the sampling decoding strategy with a temperature of 0.2, and top_p set to 0.95. The maximum generation length was capped at 2048 tokens.

A.4 SFT Training Details

For the Supervised Fine-Tuning (SFT) baseline shown in the main experiments and the task sensitivity analysis, we utilize the LLaMA-Factory (Zheng et al., 2024) for training. For the former, we use the full set of training data described in Section. The model is trained on NVIDIA H200 140G GPU for 8 epochs with a learning rate of e-5 and a batch size of 1. For the latter, which is used as a cold start initialization for the subsequent RL phase, we fine-tune the model only on the medical-o1-reasoning-SFT dataset annotated with `<think></think>` and `<answer></answer>` tags. This model is trained for 3 epoch with the same learning rate and batch size.

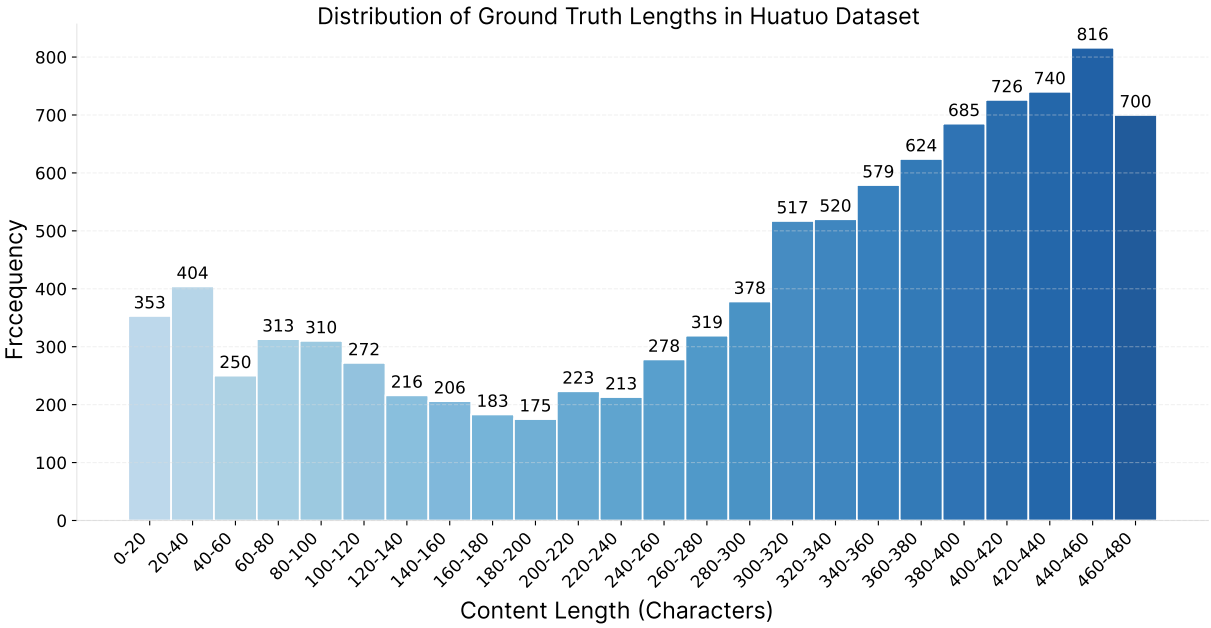

Figure 7: Character-level length distribution of all ground truth answers in the medical-o1-reasoning-SFT (Huatuo) dataset. Unlike value-based tasks that typically require only one or two words, open-ended generation involves free-form responses with varying lengths and formats. Notably, only 353 ground truth answers fall within the 0–20 character range, highlighting the general length variability of this dataset.

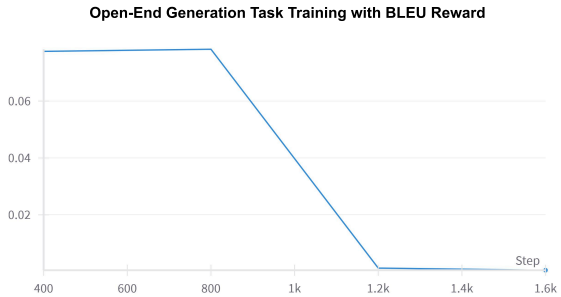

Figure 8: BLEU score as the reward signal.

A.5 Training Dynamics on Task Sensitivity Experiments.

As shown in Figure 9 (a), training with pure reinforcement learning in the MedCalc dataset leads to a more stable validation accuracy and improved performance relative to the pre-SFT-then-RL paradigm. Since MedCalc-Bench involves numerical reasoning tasks in a medical context, it requires models to possess strong computational reasoning abilities. Although SFT enriches the model with medical domain knowledge, it may interfere with the underlying general reasoning capabilities (Huang et al., 2025a; Abdin et al., 2025). Figure 9 (b) and (c) further validate this trend: SFT provides a strong initialization by enhancing the model's understanding of domain-specific terms, whereas training solely with RL leads to more stable and gradually improving performance. Thus, for tasks that involve complex numerical inference in medicine, we find pure RL to be a more appropriate training strategy.

A.6 Training Dynamics on BLEU Reward

When using BLEU as the reward signal to guide training on open-ended generation tasks, we observe catastrophic training collapse once the model fails to discover trajectories yielding higher reward.

B Exploration of Open Generation Tasks Reward.

Figure 8 shows the training dynamics. As mentioned in Failure Case of BLEU, the model struggles to consistently achieve high reward $R_{correct}$, largely due to the rigid constraints imposed by the output format, highlighting a critical need for more flexible reward that better capture the semantic quality and correctness of generated answers.

C Case Study of Different Training Paradigm.

To better understand the nuanced differences in output behaviors across training paradigms (Base, SFT, and RL) and model scales, we conduct a qualitative case study. As illustrated in Figure 11, after acquiring both background information about the

patient and the specific clinical query, the model must identify which pieces of information are relevant to solve the problem and which are distractors. In this particular case, body weight serves as the key feature, whereas indicators like blood pressure and pulse are irrelevant to the final answer. During the initial reasoning phase, all models can make a generally correct judgment. However, we observe that several models tend to second-guess the initial analysis. While this act of rethinking the initial reasoning is encouraging, it inadvertently leads to incorrect final outcomes in this case. We attribute this behavior to suboptimal sampling strategies and a lack of domain knowledge. In contrast, models trained with SFT or optimized through RL exhibit more stable and reliable reasoning paths, demonstrating improved robustness in their inferential process. Morever, in the following arithmetic reasoning, Med-U1 can arrive at correct answers more efficiently than models of comparable size, corroborating our experimental findings that reinforcement learning significantly enhances the model's reasoning and computational capabilities.

D Self-Consistency Evaluation

To evaluate the robustness of the reasoning process, we perform self-consistency (SC) inference by sampling multiple reasoning paths per query and aggregating the final predictions through majority voting. Each instance is decoded with five independent samples. The results are shown in Table 5. The SC setup leads to consistent performance improvements across most datasets, particularly for tasks requiring longer contextual reasoning, indicating that the reasoning process in Med-U1 exhibits stable convergence behavior.

The improvement under self-consistency suggests that Med-U1 benefits from multiple reasoning samples, which reduce stochastic variability and enhance reliability of the final predictions.

E Ablation on the Length Reward

To further clarify the necessity of the length reward, we trained models both with and without the *Target Length Matching Reward*. As shown in Table 6, the inclusion of this reward has minimal impact on raw accuracy but enables controllable reasoning under specified token budgets, aligning with our design goal of user-adjustable reasoning.

Table 5: Performance comparison (%) under the self-consistency (SC) setup. "+SC" denotes aggregation over five reasoning paths. Dataset abbreviations: ENQA = EHRNoteQA, MedXQA = MedXpertQA.

Model	**MEDiQ**	**ENQA**	**MedXQA**	**MedCalc**
HuatuoGPT-o1 (7B)	72.66	44.83	14.29	22.11
+ *SC*	71.32	45.98	13.06	22.86
FineMedLM-o1 (8B)	60.16	40.23	10.61	14.31
+ *SC*	62.02	41.38	10.12	14.91
ClinicalGPT-R1 (7B)	56.25	43.68	11.84	18.59
+ *SC*	56.59	45.98	11.02	19.98
M1-23K (7B)	57.03	47.13	13.06	18.69
+ *SC*	58.14	55.17	12.24	19.89
Med-U1-3B	50.78	55.17	12.90	39.07
+ *SC*	51.56	57.47	13.29	41.85
Med-U1-7B	65.63	52.87	17.74	57.55
+ *SC*	65.63	60.92	18.55	57.85

Table 6: Ablation on the length reward (LR). Models with LR support length-constrained reasoning during inference.

Dataset	**Med-U1-3B**	**Med-U1-3B w/ LR**
MEDiQ	50.78	50.00
EHRNoteQA	55.17	52.87
MedXpertQA	12.90	14.28
MedCalc	39.07	37.97

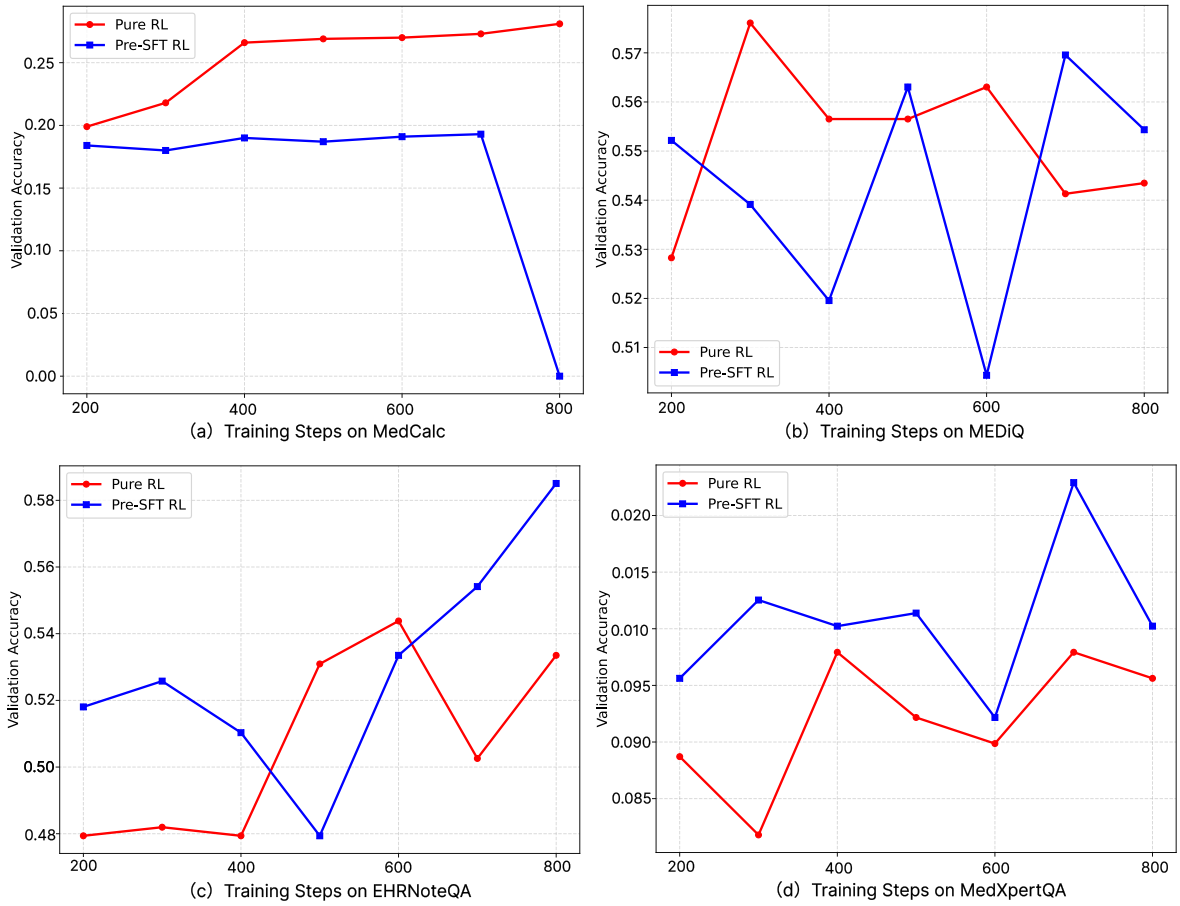

Figure 9: Training dynamics under different training strategies. The red curve denotes the pure RL paradigm, while the blue curve represents the SFT-then-RL approach.

EHRNoteQA

Problem:
What injury did the patient sustain from his mechanical fall and how was it managed during their hospital stay?

Choices:
(A) The patient broke his right rib and it was managed by bandaging the area tightly.
(B) The patient suffered from a subcapsular kidney hematoma which was managed by constant kidney flushing.
(C) The patient had a subcapsular spleen injury managed with a chest tube insertion.
(D) This patient had a small subcapsular liver hematoma, managed by following chest tube output and performing serial abdominal exams.
(E) The patient had a subcapsular liver injury, managed through a series of minor surgeries.

Answer: (D)

MedXpertQA

Problem:
An 11-year-old boy with cerebral palsy is referred to an orthopedic specialist by his pediatrician due to worsening thoracic spine curvature ..... How would the patient's current spinal curve best be described?

Choices:
(A) Moderate scoliosis with potential for neurological complications
(B) Severe scoliosis with high risk of progression
(C) Mild to moderate scoliosis requiring conservative management
(D) Progressive adolescent idiopathic scoliosis
(E) Severe scoliosis with minimal clinical significance
(F) Severe scoliosis, concerning for respiratory and cardiovascular compromise
(G) Progressive scoliosis requiring immediate surgical intervention
(H) Severe scoliosis, concerning for respiratory compromise
(I) Severe scoliosis with compensatory curve development
(J) Moderate scoliosis

Answer: (H)

MEDiQ

Problem:
A 23-year-old woman comes to the physician because she is embarrassed about the appearance of her nails. She has no history of serious illness and takes no medications. She appears well. A photograph of the nails is shown. Which of the following additional findings is most likely in this patient?

Choices:
(A) Silvery plaques on extensor surfaces
(B) Flesh-colored papules in the lumbosacral region
(C) Erosions of the dental enamel
(D) Holosystolic murmur at the left lower sternal border

Answer: (A)

MedCalc Bench

Problem:
A 65-year-old African-American man comes to the physician for a follow-up examination after presenting with elevated blood pressure readings during his last visit. He has no history of major medical illness and takes no medications. He is 180 cm (5 ft 9 in) tall and weighs 68 kg (150 lb); BMI is 22 kg/m2. His pulse is 80/min and blood pressure is 155/90 mm Hg. Laboratory studies show no abnormalities.
Based on the patient's weight, what is the patient's maintenance fluid in mL/hr?

Lower Limit: 102.6
Upper Limit: 113.4
Ground Truth Explanation:
The patient's weight is 68.0 kg. For patient's whose weight is greater than 20 kg, the formula for computing the maintenance fluid is 60 mL/hr + 1 mL/kg/hr * (weight (in kilograms) - 20 kilograms). Hence, plugging into this formula, we get 60 mL/hr + 2 mL/kg/hr * (68.0 kg - 20 kg) = 108.0 mL/hr. Hence, the patient's fluid maintenance is 108.0 mL/hr

Answer: 108.0

Medical-o1-reasoning-SFT

Problem 1:
A corrective lens has a refractive power of -10.00 diopters and is intended to correct a myopic eye. If this lens is positioned 12 mm from the eye, what is the effective power of the lens?

Answer 1:
The effective power of the lens when positioned 12 mm from the eye is approximately -8.93 diopters. This adjustment accounts for the vertex distance, allowing the lens to properly correct the myopic eye.

Problem 2:
Which single drug regimen is associated with a significant decrease in tumor size in 15-20% of patients with carcinoma esophagus?

Answer 2:
Cisplatin is a single drug regimen associated with a significant decrease in tumor size in 15-20% of patients with carcinoma of the esophagus. It is a platinum-based chemotherapeutic agent commonly used in treating various carcinomas, including esophageal cancer, and has shown efficacy as a monotherapy in shrinking tumors.

Figure 10: An overview of the five datasets used in our experiments is presented, with examples drawn from each. Datasets highlighted with yellow borders correspond to multiple choice tasks, those with red borders represent numerical reasoning tasks, and those with blue borders correspond to open-ended generation tasks. For each dataset, we provide the expected answer format as a reference, which informs the design of our rule-based reward function.

Case Study of MedCalc-Bench Analysis

<Context> A 65-year-old African-American man comes to the physician for a follow-up examination after presenting with elevated blood pressure readings during his last visit. He has no history of major medical illness and takes no medications. He is 180 cm (5 ft 9 in) tall and weighs 68 kg (150 lb); BMI is 22 kg/m2. His pulse is 80/min and blood pressure is 155/90 mm Hg. Laboratory studies show no abnormalities.

<Question> Based on the patient's weight, what is the patient's maintenance fluid in mL/hr? ,

<Lower Limit> 102.6

<Upper Limit> 113.4

<Answer> 108.0

<Ground Truth Think> The patient's weight is 68.0 kg. For patient's whose weight is greater than 20 kg, the formula for computing the maintenance fluid is 60 mL/hr + 1 mL/kg/hr * (weight (in kilograms) - 20 kilograms). Hence, plugging into this formula, we get 60 mL/hr + 2 mL/kg/hr * (68.0 kg - 20 kg) = 108.0 mL/hr. Hence, the patient's fluid maintenance is 108.0 mL/hr.",

Qwen-Instruct 3B/7B Base

<Think> The patient weighs 68 kg. The maintenance fluid formula for adults is 60 mL/hr plus 2 mL/kg/hr for each kg over 20, right? Wait, no, maybe it's 1.5 mL? Hmm I,'m a bit confused. Let me think. I think the standard formula is 60 + 2*(weight-20). So 68-20 is 48. 48*2 is 96. Add that to 60 gives 156 mL/hr. But wait, maybe I'm mixing up different formulas. I've heard of 4-2-1 rules for pediatrics, but for adults, I think it's 60 + 2*(weight-20). I'll go with 156 mL/hr.

<Answer> 156.0

<Error> Used a multiplier of 2 mL/kg/hr instead of 1 mL/kg/hr in the formula.

<Think> For calculating maintenance fluids in adults, the formula is 40 mL/hr plus 1 mL/kg/hr for each kg over 20. So the patient is 68 kg. 68 minus 20 is 48. Multiply that by 1 gives 48. Add to 40 base rate, total is 88 mL/hr. Wait, but I'm not sure if the base rate is 40 or 60. I think for pediatrics it's different, but for adults, maybe it's 40? I'll go with 88 mL/hr.

<Answer> 88.0

<Error> Used a base rate of 40 mL/hr instead of the correct 60 mL/hr.

Qwen-Instruct 3B/7B SFT

<Think> The maintenance fluid calculation is straightforward. It's 1 mL/kg/hr. So for a 68 kg patient, that would be 68 mL/hr. Wait, but I think there's a different formula for weights over 20 kg. Let me recall. Oh right, for over 20 kg, it's 60 mL/hr plus 1 mL/kg/hr. But does that mean 60 + 1*68? That would be 128 mL/hr. Or is it 60 + 1*(68-20)? That would be 60 + 48 = 108. I'm getting confused. I'll go with the simpler 1 mL/kg/hr, so 68 mL/hr.

<Answer> 68.0

<Error> Used the entire weight (68 kg) in the formula instead of (weight - 20 kg).

<Think> I remember that maintenance fluids have different calculations based on weight. Since this patient is 68 kg, which is well over 20 kg, I use the formula for >20 kg. The formula is 60 mL/hr plus 1 mL per kg per hour for each kg above 20. So 68 minus 20 is 48. 48 times 1 is 48. Adding that to 60 gives 108 mL/hr. I think that's correct. Maybe I should check if there's a different formula for older patients, but I don't recall any age-related adjustments for maintenance fluids. So I'll go with 108.

<Answer> 108.0

Med-U1(Qwen-based 3B/7B) (Ours)

<Think> The patient's weight is 68 kg. To calculate maintenance fluids, I remember that for adults over 20 kg, the formula is approximately 60 mL/hr plus 1 mL per kg per hour for each kg over 20. Let's see, 68 minus 20 is 48. Multiply that by 1 gives 48. Adding this to the base 60 mL/hr results in 108 mL/hr. I think this is within the acceptable range. Maybe I should check if there's a different formula, but I believe this is the standard calculation. So, around 108 mL/hr seems about right.

<Answer> 108.0

<Think> The patient weighs 68 kg. Using the standard maintenance fluid formula for adults over 20 kg, which is 60 mL/hr plus 1 mL per kg per hour for each kg above 20. Calculating 68 minus 20 gives 48. Multiply 48 by 1 gives 48. Adding this to the base rate of 60 mL/hr results in 108 mL/hr. I think this is within the expected range. I might have heard of different calculations, but this seems to fit the standard formula. So, approximately 108 mL/hr is the answer.

<Answer> 108.0

Figure 11: A case study comparing Qwen-Instruct-Base, Qwen-Instruct-SFT, and our Med-U1 method on MedCalc-Bench, highlighting how Med-U1 effectively addresses model hallucination and calculation inaccuracies in determining patient maintenance fluid requirements.