

Agentic Reward Modeling for Medical Reasoning

Hang Zhang¹, Ruheng Wang², Yuelyu Ji¹, Mingu Kwak¹, Xizhi Wu¹,
Chenyu Li¹, Li Zhang¹, Wenqi Shi³, Yifan Peng⁴, Yanshan Wang¹

¹University of Pittsburgh ²UT Southwestern Medical Center
³Washington University in Saint Louis ⁴Weill Cornell Medicine

Abstract

Large language models have achieved strong performance on medical reasoning benchmarks, yet their deployment in clinical settings demands rigorous verification to ensure factual accuracy. While reward models offer a scalable approach for reasoning trace verification, existing methods face two limitations: they produce only scalar reward values without explicit justification, and they rely on single-pass retrieval that precludes adaptive knowledge access as verification unfolds. We introduce Med-TIV, an agentic framework that addresses these limitations by training medical reward models to iteratively query external medical evidence during evaluation. Our approach combines tool-augmented verification with an iterative reinforcement learning paradigm that requires only trace-level supervision, alongside an adaptive curriculum mechanism that dynamically adjusts training data distribution. Across four medical reasoning benchmarks, Med-TIV achieves substantial gains over existing methods, improving MedQA accuracy by 23.5% and MedXpertQA by 32.0% relative to the base generator. Crucially, Med-TIV demonstrates an 8× reduction in sampling budget compared to prior reward model baselines. These findings establish that agentic medical reward modeling offers a principled path toward more reliable medical reasoning systems.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in medical reasoning, achieving competitive performance on medical knowledge benchmarks (Ji et al., 2025; Xiao et al., 2026; Song et al., 2026). Beyond static question answering, recent work has increasingly explored LLM-based medical agents that interact with clinical environments through tool use, multi-turn dialogue, and sequential decision-making (Schmidgall

Corresponding authors: haz269@pitt.edu, yanshan.wang@pitt.edu

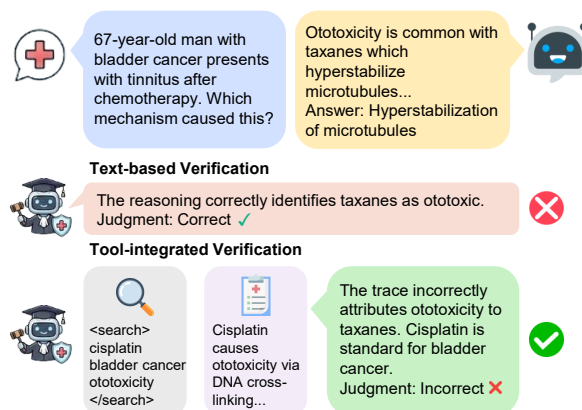


Figure 1: Comparison of medical reasoning verification paradigms. Text-based judges rely on parametric knowledge and may validate erroneous reasoning, while tool-integrated judges dynamically retrieve evidence to ground their judgments.

et al., 2024). As these systems advance toward autonomous clinical deployment, ensuring that each intermediate reasoning step is factually accurate and logically sound becomes a critical bottleneck (Zhang et al., 2026; Wang et al., 2025, 2026b). Reward-based judges offer a scalable approach to this challenge, supporting both agentic reinforcement learning and inference-time scaling through search (Snell et al., 2024). These judges range from Outcome Reward Models (ORMs) providing trace-level supervision to Process Reward Models (PRMs) offering step-level feedback, and more recently to generative judges that produce natural-language critiques (Liu et al., 2025c; Xiong et al., 2025).

Despite their effectiveness, reward-based judges exhibit fundamental limitations when applied to clinical reasoning tasks (Yun et al., 2025). A primary concern is the prevalence of hallucinations in critique traces, where judge models generate plausible yet factually incorrect assessments (Figure 1). This issue is particularly noticeable in the medical domain, where reliable verification demands

grounding in authoritative clinical evidence and established medical knowledge. Unverified judgments could lead to the propagation of incorrect diagnostic or treatment recommendations. Existing medical reasoning verifiers typically provide only scalar reward signals, offering little or no justification for their judgments and thus limiting interpretability (Jiang et al., 2025b). Furthermore, these methods often rely on a static Retrieval-Augmented Generation (RAG) pipeline, in which a fixed set of retrieved documents is prefixed to the context and remains unchanged throughout evaluation (Yun et al., 2025). Such static design precludes adaptive, multi-turn evidence gathering and forces the verifier to a fixed retrieval budget, thus limiting scalability (Wang et al., 2026a).

To address these issues, we propose Med-TIV (**M**edical **T**ool-**I**ntegrated reasoning **V**erifier), an agentic reinforcement learning framework that trains LLM-based verifiers to actively retrieve external medical evidence when judging reasoning traces¹. Med-TIV combines three key components: a tool-augmented verification paradigm for dynamic, multi-turn evidence retrieval; an iterative RL training procedure that improves verification capability using only trace-level supervision; and an adaptive curriculum strategy that matches training difficulty to the verifier’s evolving competence. Unlike static RAG pipelines that condition the model on a fixed set of retrieved documents, Med-TIV allows evidence gathering to unfold adaptively during verification, grounding judgments in external medical knowledge and improving both factual reliability and interpretability (Ji et al., 2026; Xia et al., 2025). More broadly, Med-TIV provides a tool-grounded reward model for medical reasoning agents. As medical AI systems move from static question answering toward autonomous agentic workflows, such reliable and interpretable reward signals become a key component of the training environment (Jiang et al., 2026b). In this sense, Med-TIV offers an early instantiation of an *agent-for-agent* paradigm, where an agentic verifier serves as the reward backbone for training and guiding downstream medical agents.

To verify the effectiveness of Med-TIV, we conduct extensive experiments on common medical reasoning benchmarks. Our results demonstrate that Med-TIV trains strong medical verifiers: when

guiding inference-time search for a 7B generator model, our trained verifier achieves relative improvements of 23.5% on MedQA and 32.0% on MedXpertQA compared to the generator model alone. Moreover, Med-TIV consistently outperforms existing medical reward model baselines and surpasses the performance of models that are up to $4\times$ larger in scale. Notably, Med-TIV also demonstrates an $8\times$ gain in sampling efficiency compared to prior reward-based approaches, achieving equivalent accuracy with substantially fewer sampled reasoning traces during test-time search.

Our main contributions are as follows:

- We propose Med-TIV, a framework that reframes medical reward models from passive scorers into evidence-seeking verifiers that actively retrieve medical knowledge, critique candidate reasoning traces, and produce grounded correctness judgments.
- We introduce an iterative RL paradigm with curriculum-based difficulty adaptation that progressively improves verification capabilities through self-bootstrapping, requiring only trace-level supervision rather than dense step-level expert annotations.
- We empirically characterize how agentic medical verifiers improve inference-time reasoning, showing that Med-TIV transfers across generator models, consistently benefits verifier-guided self-consistency, and changes the sampling-efficiency trade-off by matching prior reward-model performance with substantially fewer sampled traces.

2 Preliminaries

2.1 Problem Setup

We define *medical reasoning verification* as the task of assessing the correctness of a multi-step reasoning trace generated in response to a medical question. Formally, given a medical question $q \in \mathcal{Q}$ and a multi-step reasoning trace $\tau = (s_1, s_2, \dots, s_m)$ from a generator model, where s_i denotes the i -th reasoning step, a verifier model determines whether τ contains any errors. We formulate this problem as binary classification, where the verifier $V_\theta(q, \tau)$ produces a judgment $\ell \in \{0, 1\}$, where $\ell = 1$ indicates an error-free reasoning trace, and $\ell = 0$ indicates the presence of one or more errors. Unlike scalar reward models that output continuous scores, we adopt a generative judge paradigm in which the verifier produces a discrete judgment accompanied by a detailed critique trace that provides

¹Code is available at <https://github.com/PittNAIL/med-tiv>

a structured justification for the decision.

2.2 Tool-Augmented Reasoning Verifier

Following prior works (Jin et al., 2025), we extend the verifier with access to an external *search engine* $\mathcal{E}$ that retrieves top- k documents from a curated medical corpus (See §A.2 for details). Retrieved documents are appended verbatim to the verifier context. Given a verification instance (q, τ) , the verifier constructs an iterative verification trajectory. At step k , the trajectory is represented as:

$$\mathbf{t}_k = \{r_1, a_1, o_1, \dots, r_k, a_k, o_k\},$$

where r_i denotes a natural language reasoning step analyzing the medical content, a_i is a search query formulated to retrieve relevant medical knowledge, and $o_i = \mathcal{E}(a_i)$ represents the retrieved documents. The iterative verification process is defined as:

$$(r_k, a_k) \sim V_\theta(q, \tau, \mathbf{t}_{k-1}),$$

$$o_k = \mathcal{E}(a_k),$$

$$\mathbf{t}_k = \mathbf{t}_{k-1} \oplus r_k \oplus a_k \oplus o_k,$$

where $\oplus$ denotes sequence concatenation. This process continues until the verifier produces a final judgment $\ell \sim V_\theta(q, \tau, \mathbf{t}_T)$ at the terminal step T . By allowing multiple tool executions, the verifier dynamically retrieves medical knowledge as needed to verify specific claims in the reasoning trace. Table 4 in §A.4 shows the explicit instruction used in our experiments.

2.3 Test-Time Search

Test-time search strategies improve reasoning performance by leveraging reward models to evaluate and select among multiple candidate solutions (Shi et al., 2024). Given a frozen generator model π_{gen} and a question q , we first sample N independent reasoning traces:

$$\{\tau^{(j)}\}_{j=1}^N \sim \pi_{\text{gen}}(\cdot | q).$$

A trained verifier V_θ then scores each candidate trace, and the final output is selected based on these scores. Common selection strategies include Best-of- N sampling, which selects the trace with the highest score:

$$\hat{\tau} = \arg \max_{\tau^{(j)}} V_\theta(q, \tau^{(j)}),$$

and verification-based majority voting, where candidate traces are first filtered by the verifier and

the final answer is determined by consensus among verified traces. Med-TIV trains such a plug-in verifier that provides tool-grounded assessments that can be used to augment decision-making for any frozen generator model at inference time.

3 Tool-Integrated Medical Reasoning Verifier

Med-TIV is an agentic verification framework that trains models to leverage external knowledge bases for verifying whether a given medical reasoning trace contains errors. We adopt an iterative training approach based on dynamic curriculum learning, which requires no fine-grained step-level expert supervision and trains solely through multiple rounds of reinforcement learning (Figure 2). We next describe the training procedure of Med-TIV in detail.

3.1 Tool-Integrated RL with Verifiable Rewards

Data Construction. All training data across iterations is derived from the open-source Med-PRM dataset². Each original instance consists of a tuple $(q, \tau, \ell_{\text{step}}, \ell_{\text{trace}})$, where q is a medical question, τ is a multi-step reasoning trace, ℓ_{step} denotes step-level labels, and ℓ_{trace} is a trace-level correctness label.

At each training iteration, we only utilize the triplet $(q, \tau, \ell_{\text{trace}})$ with human-annotated trace-level labels. Step-level labels ℓ_{step} are intentionally excluded, as Med-TIV is designed to improve verification performance without relying on fine-grained supervision. For each training iteration, we fix the training data budget to 20K instances and enforce a balanced label distribution between correct ($\ell_{\text{trace}} = 1$) and incorrect ($\ell_{\text{trace}} = 0$) reasoning traces.

Algorithm. We employ Dr. GRPO (Liu et al., 2025b) as the RL algorithm for training the verifier. Given a verification instance (q_i, τ_i) , we sample a group of G verification trajectories $\{\mathbf{o}_i\}_{i=1}^G$ from the current policy π_θ . Each trajectory $\mathbf{o}_i = (o_i^1, \dots, o_i^{|\mathbf{o}_i|})$ consists of reasoning tokens, search queries, retrieved documents, and a final judgment. The objective is:

$$\frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[r_i^t \hat{A}_i^t, \text{clip} \left(r_i^t, 1 - \epsilon_l, 1 + \epsilon_h \right) \hat{A}_i^t \right] \right\}, \quad (1)$$

²Dataset is available at <https://huggingface.co/datasets/dmis-lab/llama-3.1-medprm-reward-training-set>

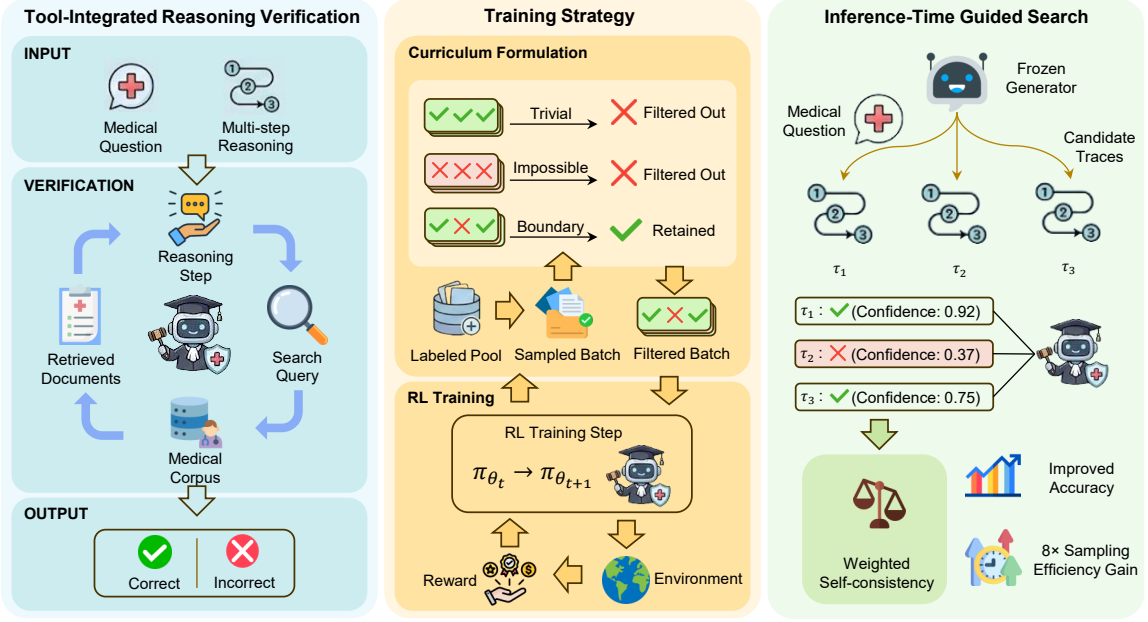


Figure 2: **Overview of Med-TIV.** *Left:* Tool-integrated verification iteratively analyzes reasoning traces, formulates search queries, and retrieves medical evidence before producing correctness judgments. *Middle:* Curriculum formulation filters trivial and impossible instances, retaining boundary cases for RL training. *Right:* At inference time, the verifier evaluates candidate medical reasoning traces generated by a frozen model and final answers are selected via weighted self-consistency.

where $r_i^t = \frac{\pi_{\theta}(o_i^t | \mathbf{q}, \mathbf{o}_i^{<t})}{\pi_{\theta_{\text{old}}}(o_i^t | \mathbf{q}, \mathbf{o}_i^{<t})}$, $\mathbf{q} = (q, \tau)$ denotes the input prompt containing the question and reasoning trace, $\mathbf{o}_i^{<t}$ represents previously generated tokens, and ϵ_l and ϵ_h are the clipping parameters. The advantage term $\hat{A}_i^t$ is defined as:

$$\hat{A}_i^t = R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\}).$$

Reward Designs. To facilitate multi-turn RL with tool execution, we design a structured reward $R = R_c \times R_f$ that combines a *correctness reward* R_c , which measures alignment between the verifier’s judgment and the ground-truth label, and a *format reward* R_f , which enforces adherence to the structured output format. The detailed reward formulation is provided in §A.5.

3.2 Training Strategies

Adaptive Curriculum Formulation. A central challenge in RL for verification is ensuring that training data remains appropriately calibrated to the evolving capability of the model. Instances that are either trivially easy or impossibly difficult yield minimal learning signal, as the resulting policy gradients approach zero. To address this issue, we adopt a model-aware curriculum formulation mechanism that dynamically adapts the task distribution at each training iteration.

Concretely, before each iteration t , we perform online filtering on the sampled batch $\mathcal{B}_t$ to construct an effective training set $\mathcal{D}_t$:

$$\mathcal{D}_t = \{(q, \tau, \ell) \in \mathcal{B}_t : \exists g, g' \in \{1, \dots, G\} \text{ s.t. } r^{(g)} \neq r^{(g')}\}.$$

Here, for each candidate instance $(q, \tau, \ell) \in \mathcal{B}_t$, we sample G verification trajectories $\{o^{(g)}\}_{g=1}^G$ from the current policy π_{θ_t} . We then compute the corresponding rewards $\{r^{(g)}\}_{g=1}^G$. Finally, we retain only instances if any two rewards are different, i.e., reward variance is non-zero (Khatri et al., 2026).

This criterion eliminates prompts where the model either consistently succeeds or consistently fails across all sampled trajectories. By filtering these zero-gradient instances, optimization is focused on decision-boundary cases where the verifier exhibits uncertainty.

To maintain a fixed training budget per iteration, we iteratively resample additional instances from the labeled pool $\mathcal{B}$ and apply the same filtering criterion until $|\mathcal{D}_t| = 20K$. This dynamic curriculum evolves naturally across iterations as the verifier improves, eliminating the need for manually designed difficulty schedules.

Iterative Training via Self-Bootstrapping. We adopt an iterative training approach that progres-

Algorithm 1 Iterative Training of Tool-Integrated Medical Reasoning Verifier

Require: Base verifier π_{θ_0} , labeled dataset pool $\mathcal{D} = \{(q_i, \tau_i, \ell_i)\}_{i=1}^N$, maximum iterations $T_{\max}$, batch size B , group size G , search engine $\mathcal{E}$

Ensure: Trained verifier π_{θ^*}

```
1: for  $t = 1$  to  $T_{\max}$  do
2:    $\triangleright$  Sample labeled batch
3:    $\mathcal{B}_t \leftarrow \text{SAMPLEBATCH}(\mathcal{D}, B)$ 
4:    $\mathcal{D}_t \leftarrow \emptyset$ 
5:    $\triangleright$  Curriculum formulation
6:   for each  $(q, \tau, \ell) \in \mathcal{B}_t$  do
7:     Sample verification trajectories:
        $\{\hat{\ell}^{(g)}\}_{g=1}^G \sim \pi_{\theta_t}(\cdot \mid q, \tau, \mathcal{E})$ 
8:     Compute rewards within group:
        $r^{(g)} \leftarrow \mathbb{1}[\hat{\ell}^{(g)} = \ell]$ , for  $g \in 1, \dots, G$ 
9:     if  $\exists g \neq g'$  such that  $r^{(g)} \neq r^{(g')}$  then
10:      Add  $(q, \tau, \ell)$  to curriculum set  $\mathcal{D}_t$ 
11:    end if
12:  end for
13:   $\triangleright$  RL optimization on curriculum data
14:   $\pi_{\theta_{t+1}} \leftarrow \text{DR.GRPO}(\pi_{\theta_t}, \mathcal{D}_t, \mathcal{E})$ 
15: end for
16: Return  $\pi_{\theta_{T_{\max}}}$ 
```

sively improves verification capabilities through multiple rounds of RL. Unlike prior work that alternates between rejection sampling, supervised fine-tuning (SFT), and RL (Xu et al., 2026), our approach operates entirely through iterative RL, following the RL-Zero paradigm where the model reinforces its verification capabilities without requiring dense turn-level expert demonstrations for cold start.

Starting from the base model π_{θ_0} , we perform $T_{\max}$ iterations. Each iteration consists of three stages:

$$\begin{aligned}\mathcal{B}_t &\leftarrow \text{SAMPLEBATCH}(\mathcal{D}, B), \\ \mathcal{D}_t &\leftarrow \text{FILTER}(\mathcal{B}_t, \pi_{\theta_t}), \\ \pi_{\theta_{t+1}} &\leftarrow \text{RL}(\pi_{\theta_t}, \mathcal{D}_t).\end{aligned}$$

Each iteration draws a fresh batch $\mathcal{B}_t$ from the annotated pool $\mathcal{D}$ with trace-level labels, ensuring a balanced distribution of correct and incorrect reasoning traces. The curriculum filtering then constructs the training set $\mathcal{D}_t$ as described above, and RL optimization updates the policy based on the structured reward. We summarize the overall training procedure in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Evaluation benchmarks. We evaluated Med-TIV on four open-source medical question-answering benchmarks: MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), MMLU-Med (Hendrycks et al., 2020), and MedExpertQA (Zuo et al., 2025), using accuracy as the evaluation metric. These benchmarks collectively assess the verifier’s ability to distinguish correct from erroneous reasoning across varying difficulty levels and medical subdomains. Detailed descriptions of benchmarks are in §C.1.

Implementation details. We trained verifiers using two light-weight backbone models: Llama3.1-8B and Qwen2.5-7B, with Llama3.1-8B as the default for results reporting. All training was conducted using the VeRL-Tool framework (Jiang et al., 2025a). Detailed hyperparameters are shown in Appendix A.1. All experiments were conducted on 4 NVIDIA H100 GPUs with 80GB of memory. Due to computational constraints, we limit the maximum number of RL iterations to $T_{\max} = 2$ and we set the group size for curriculum formulation (Section 3.2) to $G = 8$.

For inference, we used the default sampling hyperparameters for all models. In reward-guided search experiments, unless otherwise specified, we used Qwen2.5-7B as the frozen generator and sampled up to 32 candidate reasoning traces per question. We applied Hard-Weighted Self-Consistency as the default test-time search strategy. Unless otherwise specified, all reported results are averaged over three independent inference runs.

Baselines. We compared Med-TIV against two groups of baselines. 1): *Off-the-shelf LLMs*: GPT-4o-mini (Hurst et al., 2024), Gemini-2.0-Flash, DeepSeek-R1 series (Guo et al., 2025), Qwen2.5 series (Qwen, 2025), Llama3.1 (Grattafiori et al., 2024), AlphaMed (Liu et al., 2025a), UltraMedical (Zhang et al., 2024), and HuatuoGPT-o1 (Chen et al., 2024). 2): *Medical domain-specialized Reward Models*: MedS³ (Jiang et al., 2025b) and Med-PRM (Yun et al., 2025). Detailed descriptions of each reward model baseline are shown in §A.3.

Test-Time Search Strategies. We evaluated three test-time search strategies that leverage Med-TIV to improve the reasoning performance of frozen generators. Given a reasoning trace

Table 1: Main evaluation results on medical reasoning benchmarks. We report accuracy (%) on MedQA, MedMCQA, MMLU-Med, and MedXpertQA. **Bold** numbers indicate the best results among the reward model group. ✓: Verifier supports external tools for judging; ✗: Verifier does not support external tools for judging.

Baselines	✗	Train	Size	MedQA	MedMCQA	MMLU-Med	MedXpertQA	Avg.
<i>Proprietary Models</i>								
GPT-4o-mini	-	-	-	79.03	68.20	87.79	17.84	63.22
Gemini-2.0-Flash	-	-	-	87.51	72.60	92.01	20.57	68.17
<i>General Reasoning Models</i>								
DeepSeek-R1	-	-	671B	90.34	78.80	94.40	37.76	75.33
R1-Distill-Qwen	-	-	7B	24.82	36.40	47.47	7.43	29.03
R1-Distill-Llama	-	-	8B	34.96	43.60	64.19	5.35	37.03
<i>General Non-reasoning Models</i>								
Qwen2.5	-	-	32B	73.21	64.83	84.94	13.87	59.21
Qwen2.5	-	-	7B	60.96	56.56	76.96	12.15	51.66
Llama3.1	-	-	8B	70.93	61.60	78.97	13.02	56.13
<i>Medical Reasoning Models</i>								
AlphaMed	-	-	7B	71.01	61.46	81.16	19.16	58.20
UltraMedical	-	-	8B	72.66	62.60	79.61	15.25	57.53
HuatuoGPT-o1	-	-	8B	72.19	63.60	75.30	16.84	56.98
<i>Medical Reward Models</i>								
MedS ³	✗	225k	7B	64.89	58.91	80.53	12.90	54.31
Med-PRM	✓	111k	7B	69.99	62.36	80.99	13.51	56.71
Med-TIV (Ours)	✓	20k	7B	75.26	64.70	85.58	16.04	60.40

$\tau = (s_1, s_2, \dots, s_K)$ with K steps, our verifier assigns a confidence score $r_\tau \in [0, 1]$ for the entire trace, defined as the softmax probability of the 1 token over the logits of both 1 and 0 tokens.

- *Best-of-N*. Given a question q , we sampled N candidate traces $\{\tau^{(j)}\}_{j=1}^N$ from the generator and selected the trace with the highest verifier confidence score:

$$\hat{\tau} = \arg \max_{\tau^{(j)}} r_{\tau^{(j)}}.$$

- *Hard-Weighted Self-Consistency*. We first filtered traces by the verifier’s binary judgment, keeping only those labeled correct ($V_\theta(q, \tau) = 1$). Among the filtered traces, we applied majority voting to determine the final answer:

$$\hat{a} = \arg \max_a \sum_{j=1}^N \mathbb{1}[V_\theta(q, \tau^{(j)}) = 1] \cdot \mathbb{1}[\text{ans}(\tau^{(j)}) = a].$$

- *Soft-Weighted Self-Consistency*. Instead of binary filtering, we weighted each trace’s vote by the verifier’s confidence score:

$$\hat{a} = \arg \max_a \sum_{j=1}^N r_{\tau^{(j)}} \cdot \mathbb{1}[\text{ans}(\tau^{(j)}) = a].$$

4.2 Main Results

Table 1 presents the main results on four medical reasoning benchmarks. Models trained with Med-TIV consistently outperform existing baselines across all benchmarks. Specifically, under guided-search using a Med-TIV-trained verifier, Qwen2.5-7B attains accuracies of 75.26% on MedQA, 64.70% on MedMCQA, 85.58% on MMLU-Med, and 16.04% on MedXpertQA, yielding an average accuracy of 60.40%. Notably, Med-TIV enables this 7B generator to rival substantially larger models, even surpassing the base performance of Qwen2.5-32B despite using a generator that is approximately $4\times$ smaller. Compared to domain-specialized medical reasoning models of similar scale, Med-TIV outperforms HuatuoGPT-o1-8B and UltraMedical-8B by 3.07% and 2.60% on MedQA, respectively, demonstrating the effectiveness of our tool-integrated verification. Case analysis in §D further illustrates how Med-TIV identifies subtle reasoning errors.

4.3 Analysis

We conducted a series of analyses to investigate key research questions regarding our proposed framework. In addition, we provide complementary eval-

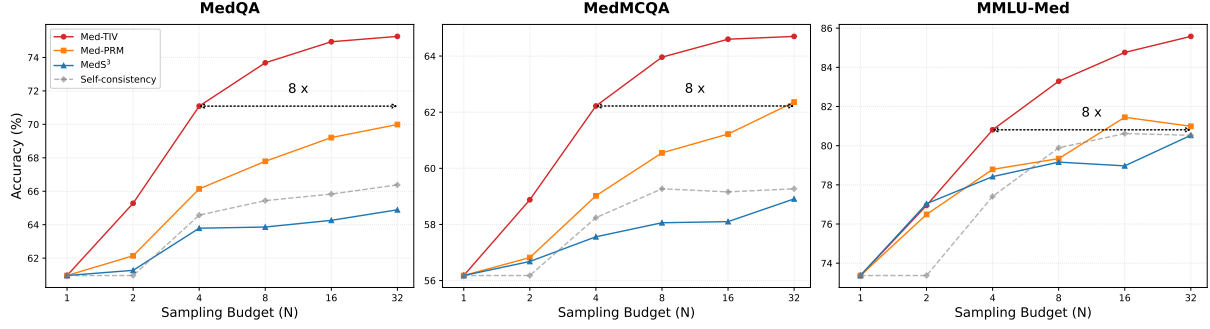


Figure 3: Test-time scaling analysis across three medical reasoning benchmarks. Each plot shows accuracy versus sampling budget $N \in \{1, 2, 4, 8, 16, 32\}$ for four baselines. Med-TIV consistently outperforms baselines across all sampling budgets and benchmarks.

uations in §B, covering direct verification classification (§B.1), open-ended clinical diagnosis (§B.2), and an ablation on RL and tool integration (§B.3).

Q1: Does Med-TIV generalize across different generator models? To evaluate the generalizability of the trained verifier, we applied Med-TIV to guide test-time search across generator models of varying sizes and capabilities. As shown in Table 2, when using Qwen2.5-7B as the generator, Hard-Weighted Self-Consistency yields a relative improvement of 23.5% over the base model’s single-sample accuracy, substantially outperforming the 12.2% gain achieved by standard Self-Consistency. Notably, the domain-specialized AlphaMed-7B model also benefits from verifier guidance with a 6.5% relative improvement, indicating that our verifier provides complementary verification capabilities beyond domain-specific fine-tuning. The improvements extend to larger models as well: Qwen2.5-32B achieves a 3.8% relative gain during test-time search, demonstrating that a light-weight 8B verifier can effectively guide models that are significantly larger than itself. This cross-scale generalization suggests that Med-TIV learns transferable verification patterns rather than overfitting to specific generator characteristics.

Q2: How do different test-time search strategies compare under Med-TIV? We then systematically compare different test-time search strategies under verifier guidance to identify the most effective approach for leveraging verification signals. As shown in Table 2, Hard-Weighted Self-Consistency consistently achieves the highest accuracy across all generators, followed by Soft-Weighted Self-Consistency and Best-of- N selection. On Qwen2.5-7B, Hard-Weighted Self-Consistency outperforms Best-of- N by 3% ab-

Table 2: Performance improvements from using Med-TIV as a verifier on MedQA. For each generator model, the first row indicates the accuracy over single sampled trace per question.

Models	MedQA
<i>Qwen2.5-7B</i>	60.96
+ Self-Consistency	66.38 (+5.42)
+ Best-of- N (Med-TIV)	72.35 (+11.39)
+ Soft Weighted SC (Med-TIV)	75.02 (+14.06)
+ Hard Weighted SC (Med-TIV)	75.26 (+14.30)
<i>AlphaMed-7B</i>	71.01
+ Self-Consistency	74.23 (+3.22)
+ Best-of- N (Med-TIV)	75.02 (+4.01)
+ Soft Weighted SC (Med-TIV)	75.33 (+4.32)
+ Hard Weighted SC (Med-TIV)	75.65 (+4.64)
<i>Qwen2.5-32B</i>	73.21
+ Self-Consistency	75.26 (+2.05)
+ Best-of- N (Med-TIV)	75.57 (+2.36)
+ Soft Weighted SC (Med-TIV)	75.96 (+2.75)
+ Hard Weighted SC (Med-TIV)	75.96 (+2.75)

solute accuracy, suggesting that majority voting among verified traces provides more robust answer selection than simply choosing the highest-confidence individual trace.

Q3: Can Med-TIV reduce the sampling budget required to achieve state-of-the-art performance compared to existing baselines? Next, we investigated how verification performance scales with sampling budget, a critical consideration for deployment under varying computational constraints. As shown in Figure 3, Med-TIV achieves substantial efficiency advantage over existing medical reward models across all three benchmarks. In particular, Med-TIV matches the performance of baselines using only 4 samples, whereas the baselines

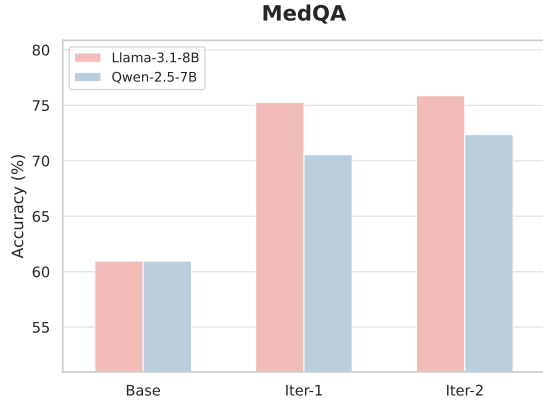


Figure 4: Ablation on base model selection and training iterations.

require 32 samples, representing an $8\times$ reduction in sampling budget. On MedQA, Med-TIV achieves 72.1% accuracy at $N = 4$, while Med-PRM requires the full $N = 32$ budget to reach 70.0% accuracy. Since the number of sampled generator traces scales approximately linearly with generator-side inference cost, this corresponds to an $8\times$ reduction in generator sampling budget. We emphasize that this is a generator-side sampling efficiency gain rather than a full end-to-end compute reduction, since verifier generation, retrieval calls, and retrieved-context processing also contribute to total system latency. We further analyze verifier-side generation length in Appendix B.5.

Q4: Does Med-TIV generalize across different base models? To assess the generality of our proposed framework, we compared verification performance using two distinct verifier backbones: Llama3.1-8B and Qwen2.5-7B. As shown in Figure 4, both backbones achieve strong performance after two training iterations. Llama3.1-8B consistently outperforms Qwen2.5-7B by approximately 3.5% absolute accuracy on MedQA, achieving 75.86% versus 72.35% after 2 iterations of training. The parallel performance gains observed across both models indicate that Med-TIV is agnostic to backbone architectures.

Q5: What is the impact of iterative training? Figure 4 presents ablation results examining the impact of iterative training with adaptive curriculum formulation. Llama3.1-8B improves from 60.96% to 75.26% after iteration 1, with marginal gains to 75.86% at iteration 2. Qwen2.5-7B follows a similar pattern, reaching 72.35% after two iterations. The rapid convergence suggests that the majority of

verification capability is acquired in the first round, with subsequent iterations refining boundary cases.

5 Related Work

Medical Reasoning Models. The application of large language models to medical reasoning has attracted considerable attention. Early efforts focused on domain-adaptive pretraining and instruction tuning on medical corpora (Wu et al., 2024; Singhal et al., 2025; Chen et al., 2023). More recent work has explored reasoning-enhanced medical models. HuatuoGPT-o1 (Chen et al., 2024) incorporates chain-of-thought reasoning with verification mechanisms, and UltraMedical (Zhang et al., 2024) combines high-quality instruction data with preference optimization. AlphaMed (Liu et al., 2025a) employs RL to improve medical reasoning capabilities. Despite these advances, most existing approaches focus on improving the generator model itself, whereas our work addresses the complementary problem of training a plug-and-play verifier that can improve any frozen generator through test-time search.

Tool-Assisted Reward and Judge Models. Standard LLM-based judges typically function as passive scorers limited by parametric knowledge. Recent work addresses this through agentic reward modeling, equipping verifiers with executable tools. Themis (Li et al., 2024) established the foundational framework by enabling access to calculators, search engines, and knowledge bases through structured tool-calling traces. TIR-Judge (Xu et al., 2026) advanced this paradigm in the general domain by integrating code execution to judge paired responses. TIM-PRM (Kuang et al., 2025) introduced independent tool queries for multi-modal verification to eliminate confirmation bias. The concept has further expanded to the Agent-as-a-Judge paradigm (You et al., 2026), which employs dynamic planning, tool augmentation and multi-agent coordination to decompose complex evaluation tasks. Our work instantiates this agentic paradigm within the medical domain, moving beyond static retrieval to iterative, evidence-grounded clinical verification.

6 Conclusion

We presented Med-TIV, an agentic RL framework for medical reward modeling. Our approach addresses key limitations of existing medical reward

models by offering explicit critique traces and enabling dynamic knowledge retrieval during verification. Empirical evaluations across four medical reasoning benchmarks demonstrate that Med-TIV substantially outperforms prior approaches. More broadly, Med-TIV introduces a general paradigm for building agentic reward systems that can be extended to other high-stakes domains requiring evidence-grounded evaluation.

Limitations

While Med-TIV demonstrates substantial improvements over existing medical reasoning verification approaches, several limitations warrant discussion and suggest directions for future research.

Process Supervision. Our current training paradigm relies solely on trace-level outcome rewards, providing no supervision on intermediate verification behaviors such as when to search, what queries to formulate, or how to integrate retrieved evidence. While this design eliminates the need for costly step-level annotations, it may lead to suboptimal search patterns or redundant retrieval operations. Future work could explore supervision for the verification task itself, or leverage techniques such as search behavior cloning from stronger models to provide denser optimization signals.

Retrieval Corpus and Tool Coverage. Med-TIV’s verification accuracy is inherently bounded by the coverage and quality of the underlying medical corpus. Our retrieval system indexes documents from PubMed abstracts and medical textbooks, which provides broad coverage of established medical knowledge but may lack recent findings, rare disease information, or region-specific clinical guidelines. Verification of reasoning traces involving cutting-edge treatments or highly specialized subspecialties may be limited by corpus gaps. In addition, the current implementation uses text retrieval as the only external tool and does not yet incorporate structured clinical decision-support tools such as drug-interaction checkers, dosage calculators, laboratory reference modules, or guideline engines. These tools can be integrated under the same action–observation interface used by Med-TIV, but evaluating such extensions requires dedicated clinical validation and is left for future work.

Language and Domain Scope. All training and evaluation are conducted on English-language med-

ical reasoning benchmarks. The generalization of Med-TIV to multilingual medical content or non-Western medical traditions remains unexplored. Additionally, while our benchmarks span multiple medical subdomains, certain specialized areas such as genomics, radiology interpretation, and surgical planning may require domain-adapted retrieval corpora for optimal verification performance.

Ethics Statement

Med-TIV is a research tool for verifying medical reasoning traces and is not intended for direct clinical decision-making. All training and evaluation data are derived from publicly available, de-identified benchmarks containing no protected health information. The use of existing artifacts is consistent with their intended purposes. While Med-TIV improves factual grounding through evidence retrieval, it remains susceptible to errors from corpus gaps or retrieval failures. Any downstream application in healthcare settings would require rigorous validation, regulatory approval, and appropriate human oversight.

Use of AI Assistants.

We used AI writing assistants only for language polishing and grammar checking. All scientific claims, experiments, analyses, and final writing decisions were made and verified by the authors.

Acknowledgement

This study was supported by the National Institutes of Health under grant number UG3TR005954, R01LM014573, and R01LM014306.

References

- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuoogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Y. Ji, W. Ma, S. Sivarajkumar, and 1 others. 2025. Mitigating the risk of health inequity exacerbated by large language models. *npj Digital Medicine*, 8:246.
- Yuelyu Ji, Hang Zhang, and Yanshan Wang. 2026. Bias evaluation and mitigation in retrieval-augmented medical question-answering systems. In *AMIA Annual Symposium Proceedings*, volume 2025, page 526.
- Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, and 1 others. 2025a. Verltool: Towards holistic agentic reinforcement learning with tool use. *arXiv preprint arXiv:2509.01055*.
- Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025b. s3: You don’t need that much data to train a search agent via rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21610–21628.
- Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2026a. Meds³: Towards medical slow thinking with self-evolved soft dual-sided process supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31319–31327.
- Tanqiu Jiang, Yuhui Wang, Jiacheng Liang, and Ting Wang. 2026b. Agentlab: Benchmarking llm agents against long-horizon attacks. *arXiv preprint arXiv:2602.16901*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.
- Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Venkata Sai Surya Subramanyam Duvvuri, Manzil Zaheer, Inderjit Dhillon, David Brandfonbrener, and Rishabh Agarwal. 2026. The art of scaling reinforcement learning compute for llms. In *International Conference on Learning Representations*, volume 2026, pages 72438–72467.
- Peng Kuang, Xiangxiang Wang, Wentao Liu, Jian Dong, and Kaidi Xu. 2025. Tim-prm: Verifying multimodal reasoning with tool-integrated prm. *arXiv preprint arXiv:2511.22998*.
- Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and 1 others. 2024. Tool-augmented reward modeling. In *International Conference on Learning Representations*, volume 2024, pages 18129–18156.
- Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025a. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025c. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Qwen. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.
- Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22294–22314.

- K. Singhal, T. Tu, J. Gottweis, and 1 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31:943–950.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Yun Song, Wenjia Zheng, Tiedan Chen, Ziyu Wang, Jiazhao Shi, and Yisong Chen. 2026. Deep neural network architectures for electrocardiogram classification: A comprehensive evaluation. *arXiv preprint arXiv:2602.17701*.
- Kanlun Wang, Zhe Fu, Wangjiaxuan Xin, Lina Zhou, and Shashi Kiran Chandrappa. 2025. Digital voices of survival: From social media disclosures to support provisions for domestic violence victims. *arXiv preprint arXiv:2509.12288*.
- Yuhui Wang, Changjiang Li, Guangke Chen, Jiacheng Liang, and Ting Wang. 2026a. Reasoning or retrieval? a study of answer attribution on large reasoning models. In *International Conference on Learning Representations*, volume 2026, pages 102235–102258.
- Yuhui Wang, Rongyi Zhu, and Ting Wang. 2026b. Self-destructive language models. In *International Conference on Learning Representations*, volume 2026, pages 15873–15899.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.
- Chengxuan Xia, Qianye Wu, Sixuan Tian, and Yilun Hao. 2025. Parallelism meets adaptiveness: Scalable documents understanding in multi-agent llm systems. *arXiv preprint arXiv:2507.17061*.
- Wanting Xiao, Junbo Jacob Lian, Kaichen Ouyang, Shaowei Gu, Zong Ke, Dedai Wei, Xinye Sha, Jiawei Wang, Shengwei Fu, Minyu Qiu, and Chaoyang Xu. 2026. [Newton downhill optimizer with application to engineering optimization and breast cancer feature selection](#). *Biomedical Signal Processing and Control*, 117:109184.
- Wei Xiong, Wenting Zhao, Weizhe Yuan, Olga Golovneva, Tong Zhang, Jason Weston, and Sainbayar Sukhbaatar. 2025. Stepwiser: Stepwise generative judges for wiser reasoning. *arXiv preprint arXiv:2508.19229*.
- Ran Xu, Jingjing Chen, Jiayu Ye, Yu Wu, Jun Yan, Carl Yang, and Hongkun Yu. 2026. Incentivizing agentic reasoning in llm judges via tool-integrated reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 156446–156468.
- Runyang You, Hongru Cai, Caiqi Zhang, Qiancheng Xu, Meng Liu, Tiezheng Yu, Yongqi Li, and Wenjie Li. 2026. Agent-as-a-judge. *arXiv preprint arXiv:2601.05111*.
- Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, and 1 others. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582.
- Hang Zhang, Qian Lou, and Yanshan Wang. 2026. Towards safe ai clinicians: A comprehensive study on large language model jailbreaking in healthcare. In *AMIA Annual Symposium Proceedings*, volume 2025, page 1519.
- 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.
- Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.
- 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 Additional Implementation Details

A.1 Hyperparameter Settings

Table 3 provides comprehensive hyperparameter configurations for Med-TIV training across both iterations. We maintain mostly consistent settings between iterations to isolate the effect of iterative training from hyperparameter tuning.

A.2 Retrieval Setup

We construct our retrieval infrastructure using a dense retrieval architecture optimized for medical domain queries. The corpus is derived from the MedRAG (Zhao et al., 2025) collection, specifically combining the PubMed and Textbooks sub-corpora into a unified index. The PubMed subset contains approximately 23.9 million biomedical abstracts covering research publications, while the Textbooks subset includes content from standard medical textbooks spanning clinical medicine, pharmacology, pathology, and related disciplines. After

Table 3: Hyperparameter configurations for Med-TIV training across iterations.

Hyperparameter	Iteration 1	Iteration 2
RL Algorithm	Dr.GRPO	Dr.GRPO
Clip ratio ϵ_l / ϵ_h	0.2 / 0.3	0.2 / 0.3
Learning rate	1×10^{-6}	1×10^{-6}
Warmup steps	10	10
Training epochs	5	5
Global batch size	256	256
Mini-batch size	256	256
Group size G	5	8
Rollout sampling temperature	1.0	1.0
Rollout top- p	0.95	0.95
Curriculum filtering	Enabled	Enabled

deduplication and quality filtering, the combined corpus contains approximately 24 million snippets.

We employ MedCPT (Jin et al., 2023) as our dense retrieval encoder, specifically the query encoder variant for encoding search queries and article encoder for encoding corpus snippets. Document embeddings are pre-computed and stored in a FAISS index using the Flat configuration for maximum retrieval accuracy, distributed across multiple GPUs using FAISS’s GPU sharding capability to enable parallel similarity search. For each search query, we retrieve the top-3 most relevant documents for both training and inference.

A.3 Baseline Setup

We describe the configuration of reward model baselines used in our experiments. For Med-PRM, which employs static retrieval-augmented generation, we equip it with the same retrieval corpus, encoder, and top-k setting as our framework to ensure a controlled comparison. MedS³ does not support external tool invocation and is therefore evaluated without retrieval augmentation. For confidence score extraction and inference hyperparameter settings, we follow the configurations specified in each baseline’s original publication.

A.4 Prompt Template

We design a structured prompt template that guides the verifier through systematic reasoning with explicit tool invocation syntax. The complete prompt is shown in Table 4.

A.5 Reward Design Details

We describe the two reward components used for multi-turn RL training of the verifier.

Table 4: Prompt template.

User Prompt:

You are a reasoning validator for medical problems. Your task is to think step by step and evaluate whether the given reasoning trace of a medical problem contains errors. First, you must always perform a step-by-step analysis to examine the entire reasoning process. Then, based on your analysis, you will make a definitive judgment.

- Use 1 if the reasoning trace is free of errors.
- Use 0 if the reasoning trace contains one or more errors.

Output Instruction:

You must conduct your step-by-step analysis inside `<think>` and `</think>` first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by `<search>` query `</search>` and it will return the top searched results between `<information>` and `</information>`. You can search as many times as you want. If you find no further external knowledge needed, you can directly provide the answer inside `<answer>` and `</answer>`, without detailed illustrations.

Medical Problem:

{The full Medical Problem.}

Reasoning Trace:

{The full Reasoning Trace.}

Correctness Reward (R_c). This component measures whether the verifier’s judgment aligns with the ground-truth label. Let $\mathbf{q} = (q, \tau)$ denote the verification prompt and $\ell \in \{0, 1\}$ the ground-truth label. We define:

$$R_c = \mathbb{1}(\text{extract}(\mathbf{o}) = \ell),$$

where $\mathbb{1}(\cdot)$ is the indicator function and $\text{extract}(\mathbf{o})$ parses the final judgment from the `<answer>` tags in the generated trajectory $\mathbf{o}$. Intuitively, $R_c = 1$ if the verifier’s decision is correct, and $R_c = 0$ otherwise.

Format Reward (R_f). To ensure reliable tool use and structured outputs, the verifier is required to adhere to a predefined format. Specifically, reasoning steps must be enclosed within `<think>` tags,

search queries within <search> tags, and the final judgment within <answer> tags. To discourage degenerate outputs, we further penalize excessive tag usage. Specifically, $R_f = 1$ if the output satisfies all formatting constraints and contains no more than 10 <answer> tag pairs; $R_f = 0.25$ if the output is correct but exhibits tag overflow; and $R_f = 0$ otherwise.

The final reward R is defined as the product of the two components:

$$R = R_c \times R_f.$$

B Additional Experiments

We present additional experiments evaluating Med-TIV beyond the closed-ended multiple-choice setting examined in the main paper. These experiments assess the generalizability of our approach across four complementary evaluation dimensions: direct verification classification in §B.1, open-ended clinical diagnosis in §B.2 and an ablation study disentangling the contributions of RL and tool integration in §B.3.

B.1 Verification as Classification

While the main experiments evaluate verifiers indirectly through their impact on test-time search accuracy, we directly assess each verifier’s ability to classify reasoning traces as correct or incorrect. For each benchmark, we use the same Qwen2.5-7B generator traces with 32 samples per question and treat verification as a binary classification task. We report four classification metrics including accuracy, error-class F1 (Err-F1), correct-class F1 (Cor-F1) and macro F1 (Mac-F1). We include Qwen2.5-7B self-verification (where the generator judges its own traces without external tools) as an additional baseline.

Table 5: Verification as classification performance averaged across four benchmarks: MedQA, MedMCQA, MMLU-Med, and MedXpertQA. **Bold** indicates the best result per column.

Verifier	Acc	Err-F1	Cor-F1	Mac-F1
MedS ³	47.30	55.72	26.33	41.02
Qwen-Self	60.79	55.09	54.35	54.70
Med-PRM	62.53	66.40	35.59	51.00
Med-TIV	71.94	66.31	66.68	66.49

As shown in Table 5, Med-TIV achieves the highest performance across all metrics, with an average

accuracy of 71.94% and macro F1 of 66.49%. A critical distinction emerges in the balance between error detection and correct-trace recognition. Med-PRM achieves a competitive Err-F1 of 66.40 but an extremely low Cor-F1 of 35.59, indicating a strong bias toward classifying all traces as erroneous; on MedQA in particular, its correct-class precision, recall, and F1 all collapse to 0%. Similarly, MedS³ exhibits a Cor-F1 of only 26.33, reflecting the same default-reject failure mode. In contrast, Med-TIV maintains balanced performance across both classes, with Err-F1 of 66.31 and Cor-F1 of 66.68, demonstrating that tool-integrated verification enables genuinely discriminative assessment rather than a default-reject strategy. The Qwen-Self baseline, despite having no access to external tools, outperforms both MedS³ and Med-PRM in macro F1, suggesting that these specialized reward models may overfit to error detection at the expense of recognizing valid reasoning.

B.2 Open-Ended Clinical Diagnosis

To evaluate Med-TIV beyond multiple-choice settings, we conduct experiments on AgentClinic-MedQA (Schmidgall et al., 2024), a benchmark of 214 open-ended clinical diagnostic cases derived from OSCE (Objective Structured Clinical Examination) scenarios. Each case presents detailed patient information including demographics, history, physical examination findings, and test results, requiring the model to produce a free-form diagnosis rather than select from predefined options.

Setup. We sample 32 reasoning traces per case from Qwen2.5-7B using a structured prompt that elicits diagnose predictions. We use best-of- N as test-time search strategy for all reward model baselines. Self-consistency serves as the RM-free baseline. Selected diagnoses are evaluated by GPT-4o, which performs one-pass semantic matching against the gold-standard diagnosis to determine correctness. To validate this automatic evaluation, one domain expert labeled 100 randomly sampled predicted-diagnosis/gold-diagnosis pairs from AgentClinic-MedQA. GPT-4o’s judgments matched the expert labels in all 100 cases. This agreement is consistent with the constrained nature of the judging task, which requires semantic-equivalence matching between predicted and gold diagnoses rather than open-ended clinical quality assessment.

As shown in Figure 5, Med-TIV achieves the

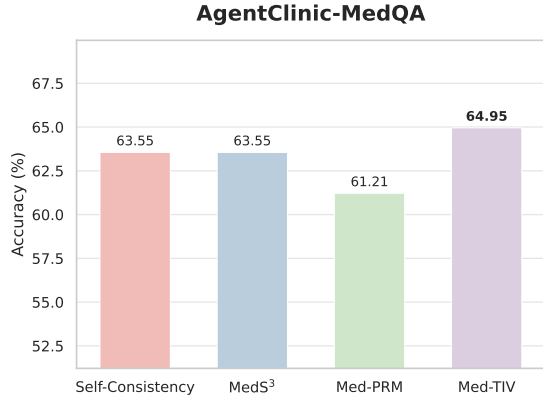


Figure 5: Open-ended diagnostic accuracy (%) on AgentClinic-MedQA.

highest accuracy of 64.95% with 139 correctly diagnosed cases, outperforming Self-Consistency at 63.55%, MedS³ at 63.55%, and Med-PRM at 61.21%. Notably, Med-PRM underperforms even the verifier-free Self-Consistency baseline, suggesting that the error-detection bias observed in §B.1 hurts performance in open-ended settings where correct diagnoses must be actively recognized rather than rejected. In contrast, Med-TIV’s balanced verification capability transfers effectively to free-form diagnostic tasks, demonstrating that the benefits of tool-integrated verification extend beyond the closed-ended multiple-choice setting examined in the main paper.

B.3 Impact of RL and Tool Integration

To disentangle the individual contributions of reinforcement learning and tool integration to Med-TIV’s overall performance,

we conducted an ablation study comparing three variants across two distinct generators: (i) the base model with no verifier guidance, (ii) Med-TIV trained with RL but without tool access (RL-only), and (iii) the full Med-TIV framework combining RL training with tool integration (RL + Tool). For the RL-only variant, the verifier is trained with

Table 6: Ablation on RL and tool integration.

Models	MedQA
Qwen2.5-7B	60.96
+ Med-TIV (RL)	69.60
+ Med-TIV (RL + Tool)	70.54
AlphaMed-7B	71.01
+ Med-TIV (RL)	76.12
+ Med-TIV (RL + Tool)	77.14

Metric	Positive rate
Query–claim alignment	100%
Evidence relevance	94%
Evidence sufficiency	79%
Critique support	77%

Table 7: Evidence-level evaluation on MedQA verification trajectories.

the same iterative Dr.GRPO procedure but the search engine is disabled during both training and inference, forcing the verifier to rely solely on parametric knowledge.

Table 6 highlights the dual benefits of our framework across two generators. RL training drives the primary gain, boosting MedQA accuracy of Qwen2.5-7B by 8.64%, confirming that the verifier effectively internalizes reasoning patterns. Tool integration provides a critical secondary boost, further elevating accuracy to 70.54%. A similar cumulative trend is observed with AlphaMed-7B. This demonstrates that while RL anchors logical verification, dynamic retrieval is essential for resolving knowledge-intensive boundary cases beyond the model’s parametric memory.

B.4 Evidence-Level Evaluation

Downstream QA accuracy alone does not directly establish whether retrieved evidence is relevant or faithfully used. We therefore conduct an evidence-level evaluation on 100 randomly selected evidence-bearing verification trajectories from MedQA. Each selected trajectory contains at least one search invocation with non-empty retrieved evidence. For each trajectory, one domain expert manually annotates four binary criteria: query–claim alignment, evidence relevance, evidence sufficiency, and critique support. These criteria evaluate whether the search query targets the disputed medical claim, whether the retrieved evidence contains information relevant to that claim, whether the evidence is sufficient to support or refute the claim, and whether the verifier’s critique is supported by the retrieved evidence.

As shown in Table 7, Med-TIV’s search queries are consistently directed at the disputed medical claims, and the retrieved evidence is usually relevant. Evidence sufficiency and critique support are lower, indicating that evidence use remains imperfect and identifying concrete directions for improving retrieval and evidence-grounded verification.

To further isolate the contribution of retrieved

Evidence condition	Accuracy
Retrieved evidence available	72.2%
Retrieved evidence masked	47.1%

Table 8: Paired evidence-masking ablation. The candidate trace, search query, and all other inputs are kept unchanged.

Verifier	Mean tokens ↓	Median tokens ↓	Std. ↓
Med-PRM	380.4	377	187.8
Med-TIV	151.6	135	95.5

Table 9: Verifier-side reasoning efficiency measured with identical tokenizer. Lower values indicate shorter generated verifier reasoning.

content, we perform a paired evidence-masking ablation. For each trajectory, we run verification under two conditions: with the retrieved observation retained and with the retrieved observation masked. The candidate trace, search query, and all other inputs are kept unchanged. As shown in Table 8, masking the retrieved evidence reduces verification accuracy from 72.2% to 47.1%, indicating that retrieved content contributes directly to the verifier’s decisions rather than merely co-occurring with successful verification.

B.5 Verifier-Side Token Efficiency

To complement the generator-side sampling analysis in Figure 3, we measure verifier-side generated reasoning length on the same set of 3,200 verification traces. We use OpenAI’s tiktoken tokenizer with cl100k_base for both Med-PRM and Med-TIV. For Med-PRM, we count the full generated evaluation output. For Med-TIV, we count only verifier reasoning tokens and exclude tokens inside <search> and <information> spans.

As shown in Table 9, Med-TIV uses 60.2% fewer verifier-generated reasoning tokens than Med-PRM on average. This analysis measures generated reasoning length rather than total end-to-end compute and does not include retrieval latency or retrieval-system overhead. Nevertheless, it shows that Med-TIV produces substantially shorter verifier-side reasoning generations than the reward-model baseline under a shared tokenizer and evaluation set.

C Benchmarks and Baselines

C.1 Benchmarks

We evaluate Med-TIV on four established closed-ended medical reasoning benchmarks that collectively assess verification capability across varying difficulty levels and medical subdomains.

- **MedQA** (Jin et al., 2021): A dataset of multiple-choice questions derived from the United States Medical Licensing Examination (USMLE), designed to evaluate clinical reasoning and medical knowledge integration across diverse specialties.
- **MedMCQA** (Pal et al., 2022): A large-scale multi-subject benchmark sourced from Indian medical entrance examinations (AIIMS and NEET-PG), covering 21 medical subjects with emphasis on factual knowledge and clinical application.
- **MMLU-Med** (Hendrycks et al., 2020): An aggregation of medical-related subsets from the Massive Multitask Language Understanding benchmark, encompassing anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine.
- **MedXpertQA** (Zuo et al., 2025): An expert-level benchmark featuring challenging questions that require multi-step clinical reasoning, differential diagnosis, and treatment planning at the level expected of practicing physicians.

C.2 Baselines

We compare Med-TIV against comprehensive baselines spanning proprietary systems, general-purpose models, and domain-specialized approaches.

Proprietary Models.

- **GPT-4o-mini** (Hurst et al., 2024): A compact variant of OpenAI’s GPT-4o optimized for efficiency while maintaining strong reasoning capabilities across diverse tasks.
- **Gemini-2.0-Flash**: Google’s efficient multi-modal model designed for fast inference with competitive performance on knowledge-intensive benchmarks.

General Reasoning Models.

- **DeepSeek-R1** (Guo et al., 2025): A 671B parameter reasoning model trained with RL, representing the current frontier of open-weight reasoning capabilities.
- **R1-Distill-Qwen / R1-Distill-Llama**: Distilled variants of DeepSeek-R1 at 7B and 8B scales

respectively, designed to transfer reasoning capabilities to smaller architectures.

General Foundation Models.

- **Qwen2.5** (Qwen, 2025): A family of open-weight language models with strong multilingual and reasoning capabilities, evaluated at 7B and 32B parameter scales.
- **Llama3.1** (Grattafiori et al., 2024): Meta’s open-source foundation model demonstrating competitive performance across diverse benchmarks, evaluated at the 8B scale.

Medical Domain Models.

- **AlphaMed** (Liu et al., 2025a): A medical reasoning model that employs RL with rule-based rewards to enhance clinical reasoning without reliance on distillation from larger models.
- **UltraMedical** (Zhang et al., 2024): A specialized medical model combining high-quality instruction tuning on curated biomedical corpora with preference optimization for improved clinical accuracy.
- **HuatuoGPT-o1** (Chen et al., 2024): A medical reasoning model incorporating chain-of-thought reasoning with internal verification mechanisms to improve diagnostic accuracy.

Medical Reward Models.

- **MedS³** (Jiang et al., 2026a): A self-evolved soft dual-sided process supervision framework for medical reasoning that generates training signals through iterative self-improvement without external annotations.
- **Med-PRM** (Yun et al., 2025): A process reward model for medical reasoning verification that provides step-level supervision using static retrieval-augmented generation with guideline-based verification.

D Case Analysis

Table 10 presents a complete verification example illustrating how a Med-TIV trained verifier identifies reasoning errors through dynamic evidence retrieval. The case involves a patient with bladder cancer who develops ototoxicity following chemotherapy. The generator’s reasoning trace incorrectly attributes the symptoms to taxanes based on their known association with ototoxicity, concluding with answer (B). However, the model retrieves evidence establishing that cisplatin—the standard neoadjuvant therapy for transitional cell

carcinoma—is the causative agent, and its mechanism involves DNA cross-linking rather than microtubule hyperstabilization. Through iterative search and reasoning, verifier correctly identifies the error, demonstrating the value of tool augmentation for catching subtle medical reasoning mistakes.

Table 10: Complete verification demonstration. Given a medical problem and a reasoning trace, the verifier retrieves relevant evidence and correctly identifies the reasoning error within the trace.

Medical Problem:

Question: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions?

Options: (A) Inhibition of proteasome (B) Hyperstabilization of microtubules (C) Generation of free radicals (D) Cross-linking of DNA

Generator Reasoning Trace:

The patient's presentation of ringing in the ear, known as tinnitus, and a sensorineural hearing loss of 45 dB, suggest ototoxicity. Ototoxicity is a well-known side effect of certain chemotherapy drugs. Neoadjuvant chemotherapy refers to the use of chemotherapy before other treatments such as surgery. The patient has received a course of neoadjuvant chemotherapy 1 week ago, which aligns with the recent onset of ototoxic symptoms. To determine the cause of ototoxicity from the provided options, we need to identify which of the options is likely to cause sensorineural hearing loss. Option (A), Inhibition of proteasome, is associated with drugs like bortezomib (Velcade), which are less commonly associated with ototoxicity. Option (B), Hyperstabilization of microtubules, is a known mechanism of taxanes (e.g., paclitaxel, docetaxel). Taxanes are known to cause ototoxicity, making this a potential candidate. Option (C), Generation of free radicals, is a general mechanism that can cause cellular damage but is not specifically known to be associated with ototoxicity. Option (D), Cross-linking of DNA, is a known mechanism of alkylating agents like cyclophosphamide. While alkylating agents can cause ototoxicity, this mechanism is not as commonly linked to ototoxicity as other options. Considering the options, the one most commonly associated with ototoxicity is hyperstabilization of microtubules. Therefore, the drug that most likely caused this patient's symptoms is one that hyperstabilizes microtubules. The answer is (B).

Verifier Judgment:

<think> Let's check the common chemotherapeutic agents used for bladder cancer and their known side effects. Transitional cell carcinoma of the bladder is often treated with cisplatin-based chemotherapy as neoadjuvant therapy. We need to determine if any of the options match the known side effects of these drugs. Let's search to confirm the side effects of cisplatin and other common bladder cancer chemotherapy drugs. </think>

<search> common side effects of cisplatin chemotherapy </search>

<information> [Search Results regarding Cisplatin toxicity and mechanisms...] </information>

<think> From the information, cisplatin is known to cause ototoxicity, which aligns with the patient's symptoms of sensorineural hearing loss. Now let's check the mechanisms of action for the options: Option (A): Inhibition of proteasome - This is associated with drugs like bortezomib and is not used for bladder cancer. Option (B): Hyperstabilization of microtubules - This is the mechanism of taxanes (e.g., paclitaxel, docetaxel). Option (C): Generation of free radicals - This is not specific to a single mechanism and is not commonly associated with ototoxicity. Option (D): Cross-linking of DNA - This is the mechanism of alkylating agents (e.g., cyclophosphamide). Cisplatin is an alkylating agent, and its mechanism involves cross-linking of DNA. Therefore, the correct answer should be (D). </think>

<answer>0</answer>