MedReflect: Teaching Medical LLMs to Self-Improve via Reflective Correction

Anonymous ACL submission

Abstract

Medical problem-solving demands expert knowledge and intricate reasoning. Recent studies of large language models (LLMs) attempt to ease this complexity by introducing external knowledge verification through retrieval-augmented generation or by training on reasoning datasets. However, these approaches suffer from drawbacks such as retrieval overhead and high annotation costs, and they heavily rely on substituted external assistants to reach limited performance in medical field. In this paper, we introduce MedReflect, a generalizable framework designed to inspire LLMs with a physician-like reflective thinking mode. MedReflect generates a single-pass reflection chain that includes initial hypothesis generation, self-questioning, self-answering and decision refinement. This self-verified and self-reflective nature releases large language model's latent capability in medical problem-solving without external retrieval or heavy annotation. We demonstrate that MedReflect enables cost-efficient medical dataset construction. With only a minimal subset of randomly sampled training examples and lightweight fine-tuning, this approach achieves notable absolute accuracy improvements across a series of medical benchmarks while significantly cutting annotation requirements. Our results provide evidence that LLMs can learn to solve specialized medical problems via self-reflection and self-improvement, reducing reliance on external supervision and extensive task-specific fine-tuning data.

1 Introduction

Recent progress of large language models (LLMs) has showcased their immense potential in medical tasks (Lin and Kuo, 2025; Xu et al., 2025). Despite this progress, deploying LLMs in professional medical scenarios offers unique challenges compared to general domains. Medical decision-making is generally narrower in scope and involves complex scenarios, which necessitate meticulous thinking to ensure more reliable answers. In addition, specialized medical terminology and intricate clinical narratives may increase hallucinated or unreliable outputs (Dou et al., 2024).

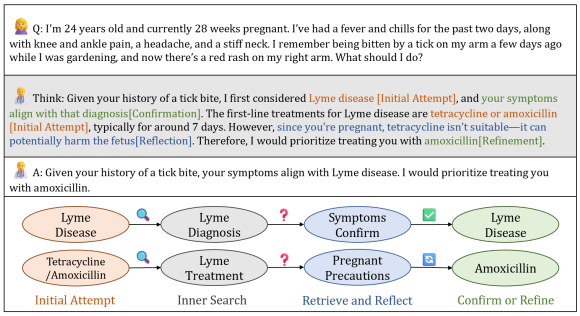

Figure 1: Example of the Physician Clinical Reasoning and Self-Correction Process Simulated by MedReflect. This figure demonstrates, through a specific clinical case, how the MedReflect framework simulates the human physician's cognitive loop of "Hypothesis-Verification-Reflection-Refinement.

Existing efforts primarily improve medical LLMs by injecting external medical knowledge. Retrieval-Augmented Generation method (RAG), for instance, consults external sources to verify and refine the generated answers (Wu et al., 2024; Lu et al., 2025). However, this method heavily relies on additional external knowledge and incurs extra storage and retrieval overhead.

Recently, reasoning techniques have been explored to guide LLMs through predefined, structured sequence of analytical steps during generation (Kwon et al., 2024). For instance, MedReason(Wu et al., 2025) leverages structured knowledge graphs to synthesize reasoning chains and construct Chain-of-Thought (CoT) data. However, these methods rely on carefully curated datasets to build reasoning paths and require domain experts to predefine comprehensive medical reasoning processes, substantially increasing annotation

costs. To alleviate this burden, HuatuoGPT-o1 (Chen et al., 2024) use an auxiliary LLM verifier to assess model-generated answers and guide the correction of errors through external feedback signals, highlighting the potential of language models to self-correct under the guidance of external signals.

However, these approaches substitute external mechanisms for the model's internal capabilities of knowledge localization and reasoning planning. This leads to a fundamental question: Can medical language models learn to generate hypotheses, retrieve relevant medical knowledge, and perform self-verification and correction—all within a single generation process? The motivation for raising this question is that LLMs already possess extensive medical knowledge during pretraining (Vladika et al., 2024). As discussed above, when faced with complex medical problems, LLMs can be guided by external verifiers to rediscover the correct reasoning path. This resembles clinical reasoning in practice: careful deliberation that involves formulating an initial hypothesis, retrieving relevant medical knowledge, and then validating or revising conclusions through iterative reasoning, which reflects the meticulous thinking required in medical decision-making. An illustrative example is shown in Figure 1.

In this work, we systematize the key components of the medical decision-making process and propose MedReflect. Our key idea is to equip the model with a directional reflection-and-correction process. We design a structured reflection mechanism that guides the model to generate and answer its own questions during reasoning, which injects direction into reflection. Through this self-questioning–self-answering procedure, the model explicitly probes uncertain assumptions, retrieves relevant medical knowledge on its own, and revises intermediate or final decisions accordingly. As a result, MedReflect does not rely on external knowledge retrieval methods, and instead, it uses self-generated questions and answers to perform internal retrieval and correction within a single generation. Experiments demonstrate that our method yields improvement on medical benchmarks. Our contributions are as follows:

- We propose **MedReflect**, a reflection-and-correction mechanism to enhance the depth and quality of reasoning in medical LLMs.
- We develop a practical approach that leverages LLMs to construct a low-cost medical reflection dataset.
- Experiments on multiple medical QA benchmarks show that MedReflect consistently improves accuracy. The results demonstrate that reflective supervision effectively teaches models to self-reflect and self-correct during generation, outperforming existing chain-of-thought training methods in both performance and training efficiency.

2 Related Works

Medical LLMs. Prior work has extensively explored developing medical-specific LLMs to excel in the medical domain (Chen et al., 2024, 2023; Team et al., 2025; Labrak et al., 2024; Zhang et al., 2024). While promising, applying LLMs to complex medical cases remains challenging (Ríos-Hoyo et al., 2024), with persistent concerns regarding ethics and hallucinations (Soffer et al., 2025). To address diagnostic capabilities, Liu et al. (2025) proposes a generalist medical LLM for diagnostic reasoning across specialties. Several approaches leverage external data or specialized training to bridge these gaps: UltraMedical (Zhang et al., 2024) comprises 410K instruction-following examples, while BioMistral (Labrak et al., 2024) utilizes PubMed Central for continued pre-training. Jeong et al. (2024) further integrates retrieval with self-reflection to enhance reliability in biomedical tasks. Recent research has increasingly focused on structured and factual reasoning. MedReason (Wu et al., 2025) leverages knowledge graphs to elicit factual reasoning steps, and MedFact-R1 (Li et al., 2025a) employs pseudo-label augmentation to bolster factual medical reasoning. Similarly, Chen et al. (2025) investigates improving syndrome differentiation thinking. To enable self-improvement, MuSeR (Zhou et al., 2025) targets medical context-awareness via multifaceted self-refinement, while HuatuoGPT-o1 (Chen et al., 2024) advances medical reasoning through a two-stage approach combining SFT, verifiable medical problems, and reinforcement learning.

Developing Models for Self-reflection and Self-correction. Self-reflection has emerged as a critical mechanism for mitigating hallucinations (Ji et al., 2023) and improving reasoning, although its effectiveness is contingent on specific conditions regarding model capability and task difficulty (Kamoi et al., 2024). Several works have studied backtracking and search as forms of self-correction (Ye et al.,

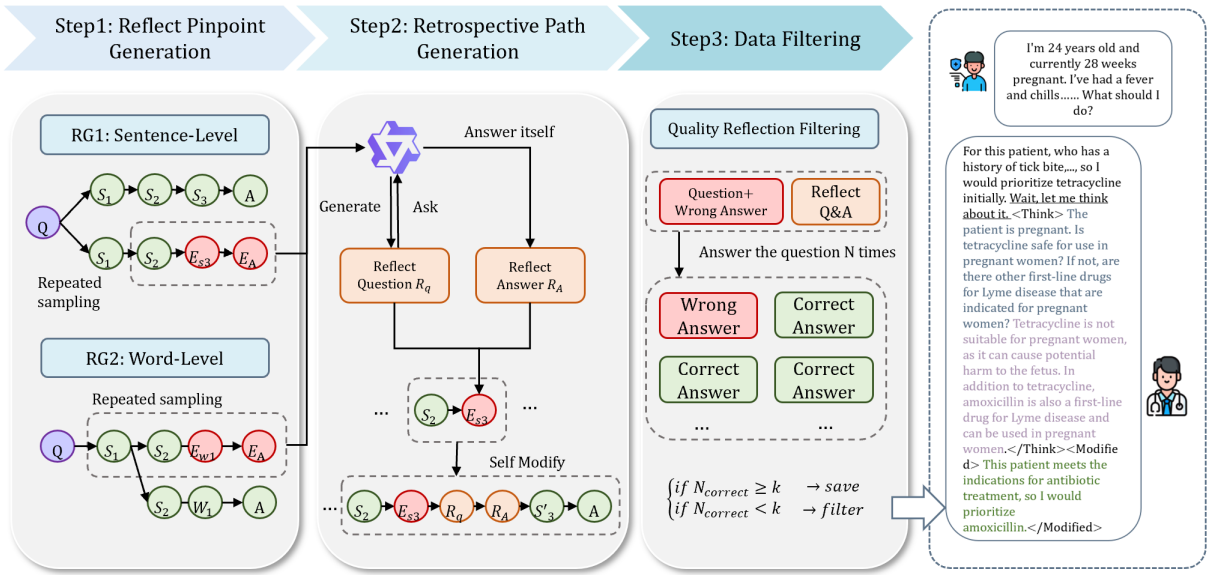

Figure 2: Data Construction Framework

2024; Qin et al., 2025). Notably, the Stream of Search (SoS) framework (Gandhi et al., 2024) enables models to self-correct by searching within language without external components, a phenomenon also explored in recent work on the emergence of thinking processes in LLMs (Ye et al., 2025). Beyond search, critique-based methods have gained traction. Xi et al. (2024) trains an expert critique model using step-level feedback to supervise the reasoning process. Building on this, ReflectEvo (Li et al., 2025b) explores iterative self-reflection to improve meta-introspection, while SaySelf (Xu et al., 2024) teaches LLMs to express confidence calibration within their self-reflective rationales. Additionally, Qu et al. (2024) introduces iterative fine-tuning to alter responses after unsuccessful attempts. More recently, Zhu et al. (2025) has begun to probe the underlying mechanisms for controlling and modulating these self-reflection behaviors.

3 Problem Setup and Preliminaries

We aim to explore whether medical LLMs can generate hypotheses, retrieve, perform self-verify and correction within a single generation process. To achieve this, we introduce MedReflect, focused on teaching LLMs to learn the reflect mode instead of directly injecting knowledge or improving their reasoning skills. The overview of the data generation of MedReflect is shown in Figure 2.

Given a medical question Q and its corresponding response trajectory $T = [S_1, S_2, \ldots, S_n, A]$, the LLM is prompted to regenerate either a selected step S_i or the entire answer. This regeneration may introduce errors, resulting in an erroneous trajectory or answer to the whole question $T_{er} = [S_1, S_2, \ldots, E_i, \ldots, E_A]$. Guided by the error E_i and the erroneous trajectory T_{er}, LLM then generates a targeted reflection question R_{q_i} along with its answer R_{a_i}. Finally, the original question Q, the erroneous trajectory T_{er}, and the reflection pair (R_{q_i}, R_{a_i}) are provided to LLM to facilitate error correction. Upon successful correction, the original correct step S_i and the reflection pair are incorporated into the trajectory, yielding the final reflective trajectory $T_{reflect} = [S_1, S_2, \ldots, E_i, R_{q_i}, R_{a_i}, S_i, \ldots, A]$.

4 Methodology

4.1 Datasource and Construction Model

We utilize two publicly available medical datasets: ChatDoctor (Li et al., 2023) and training split of MedMCQA (Pal et al., 2022). ChatDoctor contains 100k real-world conversations between patients and doctors sourced from HealthCareMagic.com. MedMCQA is a comprehensive multiple-choice question-answering dataset created specifically from real medical entrance exam questions. We used Qwen2.5-32B-Instruct to complete the entire data construction process.

4.2 Reflect Pinpoint Generation

To better control the diversity of generated pinpoints, we designed multiple reflective pinpoint generation pathways(RG), including reflective pro-

	Sentence-Level(MedMCQA)	Word-Level(ChatDoctor)
Question	A 60 yr old chronic smoker presents with painless gross hematuria of 1 day duration. Which is the investigation .....? A: USG, B: X-ray KUB, C: Urine routine, D: Urine microscopy for malignant cytology cells	I'm experiencing a throbbing ache and severe fatigue behind my right knee, ... Any ideas what might be causing this?
Generate Details (🤔: pinpoint we insert)	LLM Re-Answer: We need to consider the most appropriate and comprehensive diagnostic tool. Ultrasonography (USG) can provide an overview of the urinary system, including the kidneys, bladder, and ureters. My answer is **OptionA[truth: Option D]**	Mask Medical Entities : Throbbing knee pain can be caused by several issues, such as a ligament tear from a sudden twist **[entity1: etiology]** ..... To determine the exact cause, you should get MRI **[entity2: medical examination and testing]** and consult an professional.
	Compare and Find Pinpoint: We need to consider the most appropriate and comprehensive diagnostic tool. Ultrasonography (USG) can provide an overview of the urinary system, including the kidneys, bladder, and ureters. 🤔.....	Retry and Find Pinpoint: Throbbing knee pain can be caused by several issues, such as a ligament tear from a **[meniscus injury❌]** 🤔 ..... To determine the exact cause, you should get **[MRI scan ✅]** and consult an professional.

Figure 3: Examples of Pinpoint Generation

cesses functioning at two different levels of detail. Figure 3 shows examples of two different levels.

RG1: Sentence Level. This approach primarily focuses on the overall rewriting of reasoning sentences generated by the LLM. We utilize this method to construct training data derived from the multi-choice MedMCQA dataset. Specifically, we perform repeated sampling by prompting the LLM to respond to the original question Q multiple times, generating new answers that include their reasoning processes. For each sampled response, we employ heuristic pattern matching to extract the multi-choice decision statement and compare it with the ground truth. If the extracted decision answer A' is incorrect, we locate the specific sentence within the response that led to the wrong option, designating it as a reflection pinpoint E_{si}.

RG2: Word Level This approach focuses on medical texts rich in entities. We used this method to construct the data of the consultation dataset ChatDoctor. We extract entity W_i from medical consultation reasoning sentence S_i and mask it with its corresponding entity type, such as diseases, etiology, and treatment. The LLM is prompted to predict and fill in the masked entity multiple times. If an attempt by the LLM is assessed as incorrect and such errors occur frequently, the sentence containing this entity serves as our reflection pinpoint E_{wi}. In contrast, responses that are assessed as correct will not be treated as such pinpoints. At this level of detail, a single physician's response can produce one to three pinpoints for reflection.

4.2.1 Retrospective Path Generation

Reflect QA Generation To create reflection questions R_q and their matching answers R_a, we provide LLM with its original answer and emphasize that its answer was wrong. Through targeted prompting, we guide LLM to formulate a focused reflective question R_q. The goal of this question is to help LLM identify the key knowledge components needed to develop a better solution.

Afterward, the LLM is then instructed to answer the reflection question using only its own knowledge, without referring to the original question or adding extra details. The purpose of this method is to engage and utilize LLM's internal knowledge, promoting thorough self-awareness and reflection.

Modification Based on Reflection Using Q, R_q and R_a as input, we guide the LLM to generate the modified statement M.

(1) For RG1, LLM regenerates sentences S_i' to replace original sentences with expression flaws, performing adaptive optimization to resolve any contextual coherence issues caused by the new sentences.

(2) For RG2, LLM tries to use the new response word W_i' to substitute inaccurate vocabulary.

These modified parts will be inspected in the next step to ensure they are qualified.

4.2.2 Data Filtering

To guarantee data quality, we first filtered out samples with insufficient reasoning or irrelevant content. We then implemented a secondary validation loop to mitigate confirmation bias by feeding the generated reflections (R_q, R_a) and the initial erroneous trajectory back into the model. A reflection

was deemed valid only if it guided the model to the objective ground truth—specifically, selecting the correct option for RG1 or restoring the masked entity for RG2. Finally, to exclude stochastic reasoning, we enforced a robustness constraint based on repeated trials. We conducted $[k]$ independent inference trials for each validated sample and calculated the success rate. Only instances exceeding a consistency threshold of $\tau = 0.8$ were retained This yielded a final dataset of 36,413 medical consultation records and 21,107 multiple-choice questions.

4.3 Training Strategy

Following the construction of the generated reflection dataset, we investigate strategies to leverage these samples to enhance the model's reasoning capabilities. We formalize the medical reasoning process as a sequential decision-making problem within a discrete semantic space. In conventional foundation models, the policy is often dominated by next-token prediction objectives derived from pre-training corpus statistics. This approach frequently results in logical disconnects or hallucinations when addressing complex medical cases. As illustrated in previous works, merely extending the Chain of Thought (CoT) is analogous to performing a directionless random walk on a semantic manifold; it induces divergent thinking but relies on stochastic retries to identify a superior solution.

To address this limitation, the **MedReflect** framework introduces a reflection operator designed to provide directed self-correction to the otherwise undirected reasoning process. This is formally defined as:

$$\tau_{reflect} = \mathcal{R}(\tau_{err}, \mathcal{K}_{int}) \Rightarrow (R_q, R_a) \quad (1)$$

Here, τ_{err} represents the initial trajectory containing error breakpoints, and $\mathcal{K}_{int}$ denotes the model's implicit medical knowledge. The reflection pair (R_q, R_a) serves as a critical control signal to rectify τ_{err}. Specifically, R_q facilitates *error attribution* by compelling the model to pinpoint knowledge deficits, while R_a leverages $\mathcal{K}_{int}$ to generate a rectified factual basis.

To instantiate this theoretical operator within the language model, we structure the reasoning process as a token-augmented trajectory. We introduce four special tokens—<Think>, </Think>, <Modified>, and </Modified>—to explicitly delineate the semantic boundaries of the reflection process. Consequently, the training instance is transformed from a standard input-output pair into a *reflective sequence* $\mathbf{y}_{seq}$, constructed as:

$$\mathbf{y}_{seq} = [\tau_{err}, \texttt{<Think>}, R_q, R_a, \texttt{</Think>}, \texttt{<Modified>}, \tau_{corrected}, \texttt{</Modified>}] \quad (2)$$

This structure forces the model to treat reflection not as an external module, but as an intrinsic part of its generative policy.

Finally, we align the reflective medical LLM with this structured reasoning pattern through supervised fine-tuning (SFT). We interpret SFT here as behavioral cloning of the reflection operator. Given the dataset $\mathcal{D}_{\text{med}} = \{(\mathbf{x}, \mathbf{y}_{seq})\}$, the optimization objective is to maximize the likelihood of the augmented reflective sequence. The loss function for the model π_θ is defined as:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}_{seq}) \sim \mathcal{D}_{\text{med}}} \left[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid \mathbf{x}, y_{<t}) \right] \quad (3)$$

By minimizing this loss, the model learns to sequentially generate the error analysis and correction steps before outputting the final answer, effectively internalizing the $\tau_{reflect}$ operator into its parameters.

5 Experiments

5.1 Benchmark

We evaluate on standard medical benchmarks: MedQA(USMLE test set) (Jin et al., 2021), MedMCQA(validation set), and PubMedQA(test set) (Jin et al., 2019). To thoroughly assess the model's abilities, we also included medical sections from multi-domain evaluation frameworks, specifically the Health and Biology topics in MMLU-Pro (Wang et al., 2024), as well as the Genetics and Molecular Biology areas in GPQA (Rein et al., 2024). These benchmarks cover tasks in various aspects such as diagnosis, treatment, knowledge, and medical reasoning.

5.2 Baselines and Models

We fine-tuned Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct using 2k training examples each, and Qwen2.5-32B-Instruct using 30k examples; all training sets were randomly sampled from the MedReflect dataset. To strictly evaluate the efficacy of our approach, we benchmarked MedReflect against a comprehensive suite of baselines, ranging from standard open-source medical and general

LLMs to closed-source models. We also explicitly compared our model against reasoning-oriented baselines (e.g., HuatuoGPT-o1) under comparable settings.

5.3 Implementation Details

We fine-tuned for 3 epochs with $lr = 1e-4$. Additionally, we applied Low-Rank Adaptation (LoRA) (Shen et al., 2022) with $\alpha = 8$ to fine-tune the LLM. The experiments were carried out on 4 NVIDIA A800 GPUs. More implementation details of training can be found in appendix C.

5.4 Main Results

We conducted a comprehensive evaluation of open-source LLMs across diverse medical benchmarks, with detailed results presented in Table 1. The experimental findings indicate that MedReflect outperforms comparable open-source models on the majority of evaluated benchmarks, establishing new state-of-the-art results for models of equivalent scale, particularly in reasoning-intensive tasks.

MedReflect Demonstrates Robust Reasoning Capabilities MedReflect-7B (Qwen) achieves significant improvements over its base model and outperforms other specialized models on key benchmarks such as MedQA and MedMCQA. Notably, it exhibits superior performance on complex reasoning benchmarks. On MMLU-Pro (Health/Biology) and GPQA, MedReflect-7B surpasses all other models in its size category. MedReflect demonstrates a more balanced and robust performance across both knowledge retrieval and complex clinical reasoning tasks.

Scalability and Competitiveness with Proprietary Models Scaling up to 32B parameters, MedReflect shows outstanding performance, surpassing significantly larger open-source models. Crucially, MedReflect-32B remains highly competitive with leading proprietary models on specific tasks. While a performance gap remains on MedQA compared to the strongest proprietary systems, MedReflect-32B significantly bridges the gap between open-source and commercial SOTA models (Achiam et al., 2023; Team et al., 2024; Guo et al., 2025).

Generalizability Across Model Architectures To verify the universality of our approach, we applied the reflection mechanism to the Llama-3.1-8B-Instruct architecture. The results show that MedReflect-8B (Llama) achieves a substantial gain over its base model, confirming that the reflection mechanism is model-agnostic. However, consistent with the base models' capabilities, the Qwen-based implementation (MedReflect-7B) retains an overall performance edge over the Llama-based variant.

The Value of Reflection These advancements underscore the efficacy of the reflection chain. MedReflect achieves these results using efficient supervised fine-tuning onine-tuned a curated dataset, without relying on the massive computational resources typically required for pre-training or reinforcement learning at the scale of 70B+ models. This suggests that the self-reflection mechanism efficiently unlocks the latent reasoning capabilities of LLMs in the medical domain.

5.5 Ablation Study

We conducted further ablation studies on MedReflect-7B across the MedQA, PubMedQA, MMLU, and GPQA datasets to investigate how the reflection mechanism and training methodologies specifically affect performance. All variants are built upon the Qwen2.5-7B-Instruct backbone and were fine-tuned on the same fixed set of 2,000 samples. These samples were randomly selected from the MedReflect dataset and subsequently reformatted for training. This controlled setup ensures that any observed performance differences stem solely from the architectural modifications to the reasoning chain. As shown in Table 2, the results support the following analyses:

Directional reflection significantly outperforms blind correction We compared MedReflect with a fine-tuning strategy that excludes the reflection step (SFT w/o Reflect). It removes the intermediate reflection process, essentially forcing the model to learn to jump directly from an error to the correct answer.

Experimental results demonstrate that MedReflect exhibits significant advantages across all benchmarks. The **SFT w/o Reflect** baseline resembles a form of random retry within the semantic space; while the model observes the correct outcome, it lacks the logical pathway to deduce the correct state from the erroneous one. While **SFT w/o Reflect** enhances performance through domain knowledge injection and direct answer supervision, MedReflect achieves superior results by equipping the model with an autonomous error-correction mechanism during the reasoning

Model	MedQA	MedMCQA	PubMedQA	MMLU-Pro		GPQA	
				Health	Biology	Genetics	Molecular Biology
<=8B Language Models							
BioMistral-7B	45.0	40.2	66.9	27.4	49.2	28.6	38.5
UltraMedical-8B	71.1	58.3	77.4	55.1	66.7	41.2	48.4
Qwen2.5-7B-Instruct	57.0	55.6	55.6	50.6	70.2	36.2	49.7
LLaMA-3.1-8B-Instruct	58.7	56.0	75.2	52.7	64.6	33.8	46.8
HuatuoGPT-o1-8B	72.6	60.4	**79.2**	58.7	68.2	48.8	59.7
MedReflect-8B(LLaMa)	64.9	72.4	75.3	53.1	64.2	56.7	60.1
MedReflect-7B(Qwen)	**75.5**	**77.1**	75.3	**62.8**	**75.8**	**65.0**	**60.3**
> 8B Language Models							
Deepseek-R1	90.1	78.8	77.2	79.2	90.8	65.0	75.3
Gemini2.5-Flash	92.0	79.7	76.2	–	98.6	–	–
Gemini2.5-Pro	92.6	71.1	75.8	–	98.6	–	–
GPT-4o	86.5	76.1	78.4	–	98.4	–	–
GPT-o3	93.3	83.3	80.0	–	98.6	–	–
UltraMedical-70B	82.2	71.8	78.4	64.8	71.1	33.8	62.9
OpenBioLLM-70B	78.3	74.0	79.0	–	**93.8**	–	–
Qwen2.5-72B-Instruct	72.7	66.2	71.7	65.3	78.8	41.2	56.8
QwQ-32B-Preview	72.3	65.6	73.7	62.0	78.1	37.5	64.5
HuatuoGPT-o1-70B	83.3	73.6	80.6	71.0	82.8	56.2	66.5
MedReflect-32B(Qwen)	**83.5**	**80.1**	**82.3**	**78**	90.8	**68.3**	**70.6**

Table 1: Model performance on biomedical QA benchmarks.

Model	MedQA	PubMedQA	MMLU-Pro		GPQA	
			Health	Biology	Genetics	Molecular Biology
Qwen2.5-7B-Instruct	57.0	55.6	50.6	70.2	36.2	49.7
SFT w/o Reflect	57.3	63.3	54.3	72.8	48.3	55.5
SFT w/o Reflect Question	61.8	67.3	63.2	73.1	55.0	58.4
SFT w/o Reflect Answer	63.7	68.3	57.9	73.3	55.0	57.3
MedReflect-7B	**75.5**	**75.3**	**62.8**	**75.8**	**65.0**	**60.3**

Table 2: Performance comparison across biomedical QA datasets using various training strategies.

process. Furthermore, MedReflect does not simply memorize correction data but rather employs a diagnostic thinking pattern analogous to that of human physicians.

The integrity of the reflection chain is crucial To deconstruct the functional necessity of different components within the reflection chain, we compared three training configurations, all deriving from the same 2,000-sample dataset:

(1) **SFT w/o Reflect Answer**: Retains only the reflection questions (R_q), removing the detailed reflective answers.

(2) **SFT w/o Reflect Question**: Retains only the reflection answers (R_a), removing the guiding self-inquiry.

(3) **MedReflect**: Includes the full pair of reflection questions and answers (R_q, R_a).

The experimental data indicates that the absence of any single component leads to performance degradation, although configurations retaining partial reflection chains still outperform the baseline without reflection (**SFT w/o Reflect**). The reflection question (R_q) serves the role of error attribution, explicitly guiding attention toward potential defects in the reasoning chain and setting a clear direction for correction. The reflection answer (R_a) serves the role of knowledge retrieval and consolidation, utilizing a declarative tone to invoke the internal knowledge base and provide a factual basis for the correction. Therefore, the complete MedReflect mechanism constructs a closed cognitive loop, where the combination of R_q and R_a plays a critical role in effectively enhancing the model's reasoning capabilities.

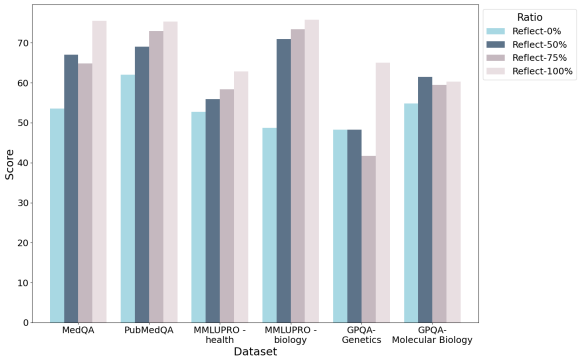

Figure 4: Results of the Reflection Data Proportion Experiment

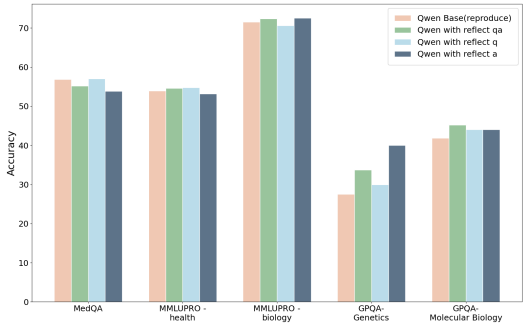

Figure 5: Results of the Attribution Analysis Experiment

5.6 Analysis

5.6.1 Data Efficiency and Marginal Utility

To investigate the marginal utility of reflection data during training, we conducted experiments with four different proportions of reflection data (0%, 50%, 75%, and 100%) while maintaining a constant total volume of 2,000 training samples. As shown in Figure 4, the results reveal a significant positive correlation between the proportion of reflection data and model performance.

Significant Marginal Returns of Reflection Data As the proportion of reflection data in the training set increases, the model's accuracy across the vast majority of medical and biological reasoning tasks demonstrates a steady upward trend. Although minor fluctuations are observed in isolated datasets, the overall trajectory indicates that a high proportion of reflection data effectively activates the model's deep reasoning capabilities. Notably, this benefit is maximized when the reflection proportion reaches 100%.

Efficiency Analysis Under the constraint of a constant total sample size, we progressively replaced standard Correct Data with Reflection Data containing detailed reasoning processes. The results show that this substitution yields significant performance gains. This implies that, per unit data sample, data containing reflection processes possesses higher information density and training efficiency. By explicitly demonstrating error correction and logical deduction, reflection data compels the model to learn reflective thinking rather than mere rote memorization.

5.6.2 Effectiveness of Reflection Content

To study the effectiveness of the reflective content produced during MedReflect generation, we collected the reflection questions R_Q, reflection answers R_A, and reflection pairs R_{QA} generated by the MedReflect-7B model, and provided them as additional context to the base Qwen2.5-7B-Instruct model. The results are shown in Figure 5, and additional experimental details are provided in Appendix C.5. As illustrated in Figure 5, overall, augmenting the original Qwen2.5-7B-Instruct model with reflection information consistently obtains better performance than the base model alone, indicating that the reflections produced by MedReflect are beneficial. On highly specialized datasets such as GPQA, both reflection questions and reflection answers lead to substantial improvements. Moreover, using only reflection questions (without answers) still brings consistent gains, suggesting that MedReflect's reflection questions help steer the direction of reasoning.

6 Conclusion

In this study, we propose MedReflect, a framework designed to enable LLMs to autonomously perform reflection and revision during medical tasks. By leveraging a lightweight LLM to construct a low-cost, diverse reflection training dataset, we trained models to acquire physician-like reflective thinking characterized. This training paradigm itself exhibits conspicuous cost-effectiveness. Notably, fine-grained reflection design proves crucial for fully activating the model's reflective capabilities. Finally, we demonstrate that MedReflect significantly enhances model performance across multiple medical QA benchmarks, robustly validating the efficacy of the reflection mechanism in improving LLM accuracy in medical tasks.

Limitation

While MedReflect significantly enhances diagnostic accuracy by mimicking the physician's reflective thinking process, this performance gain comes with an inherent trade-off in inference efficiency. As described in our method, the framework explicitly inserts a reflection chain before generating the final modified decision. Consequently, this mechanism inevitably increases the total number of generated tokens per query compared to direct-answer models. The increased sequence length results in higher computational costs and longer inference latency, which may pose challenges for deployment in real-time or resource-constrained medical consultation scenarios. Future work could explore methods to trigger reflection adaptively only when the model detects high uncertainty, thereby balancing accuracy and efficiency.

Ethics Statement

MedReflect is built upon the Qwen2.5-Instruct architecture and, like all Large Language Models (LLMs), inherits intrinsic limitations such as the susceptibility to hallucinations and the potential generation of counterfactual medical advice. In high-stakes medical decision-making scenarios, any inaccuracies or misinterpretations of clinical narratives could lead to adverse outcomes.

Therefore, MedReflect is intended solely for research purposes to explore the potential of autonomous reflective thinking in medical AI. It is not designed for direct clinical deployment without expert human oversight. Researchers and developers utilizing this framework must be cognizant of these risks and implement robust safeguards, such as rigorous secondary validation loops. We further declare that the training data derived from public datasets (ChatDoctor and MedMCQA) was processed strictly for academic research, adhering to the data filtering protocols described in this work to ensure quality and privacy compliance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

C. Chen, X. Wang, and M. Guan. 2025. Evaluating and improving syndrome differentiation thinking ability in llms. *arXiv preprint*.

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

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, and 1 others. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, and Yu Huang. 2024. Detection, diagnosis, and explanation: A benchmark for Chinese medial hallucination evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4784–4794, Torino, Italia. ELRA and ICCL.

Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D Goodman. 2024. Stream of search (sos): Learning to search in language. *CoRR*.

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*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical

reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18417–18425.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Gengliang Li, Rongyu Chen, Bin Li, Linlin Yang, and Guodong Ding. 2025a. Medfact-r1: Towards factual medical reasoning via pseudo-label augmentation. *arXiv preprint arXiv:2509.15154*.

Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. 2025b. Reflectevo: Improving meta introspection of small llms by learning self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16948–16966.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Chihung Lin and Chang-Fu Kuo. 2025. Roles and potential of large language models in healthcare: A comprehensive review. *Biomedical Journal*, page 100868.

X. Liu and 1 others. 2025. Medical large language model for diagnostic reasoning across specialties. *arXiv preprint*.

Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Sin Yee Goi, and Jinzhuo Wang. 2025. Doctorrag: Medical rag fusing knowledge with patient analogy through textual gradients. *Preprint*, arXiv:2505.19538.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Tian Qin, David Alvarez-Melis, Samy Jelassi, and Eran Malach. 2025. To backtrack or not to backtrack: When sequential search limits model reasoning. *arXiv preprint arXiv:2504.07052*.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2024. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

A. Ríos-Hoyo, N.L. Shan, and A. Li. 2024. Evaluation of llms as a diagnostic aid for complex medical cases. *arXiv preprint arXiv:2410.xxxxx*.

Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and 1 others. 2022. Lora: Low-rank adaptation of large language models.

S. Soffer, V. Sorin, and G.N. Nadkarni. 2025. Pitfalls of large language models in medical ethics reasoning. *arXiv preprint*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Medreqal: Examining medical knowledge recall of large language models via question answering. *arXiv preprint arXiv:2406.05845*.

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. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Juncheng Wu, Wenlong Deng, Xingxuan Li, and 1 others. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *Preprint*, arXiv:2408.04187.

Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Dou, Wenyu Zhan, and 1 others. 2024. Enhancing llm reasoning via critique models with test-time and training-time supervision. *CoRR*.

Dexuan Xu, Yanyuan Chen, Zhongyan Chai, Yifan Xiao, Yandong Yan, Weiping Ding, Hanpin Wang, Zhi Jin, Wenpin Jiao, Weihua Yue, and 1 others. 2025. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, page 103455.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. *arXiv preprint arXiv:2405.20974*.

Guanghao Ye, Khiem Duc Pham, Xinzhi Zhang, Sivakanth Gopi, Baolin Peng, Beibin Li, Janardhan Kulkarni, and Huseyin A Inan. 2025. On the emergence of thinking in llms i: Searching for the right intuition. *arXiv preprint arXiv:2502.06773*.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems. *arXiv preprint arXiv:2408.16293*.

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.

Yuxuan Zhou, Yubin Wang, Bin Wang, Chen Ning, Xien Liu, Ji Wu, and Jianye Hao. 2025. Enhancing the medical context-awareness ability of llms via multifaceted self-refinement learning. *arXiv preprint arXiv:2511.10067*.

Xudong Zhu, Jiachen Jiang, Mohammad Mahdi Khalili, and Zhihui Zhu. 2025. From emergence to control: Probing and modulating self-reflection in language models. *arXiv preprint arXiv:2506.12217*.

A Implementation Details for Data Generation

As outlined in the Methods section, we utilized the MedMCQA dataset to construct reflections at the sentence level and the ChatDoctor dataset to build reflections at the word level. The implementation details of this process are described below, with all corresponding prompts provided in Table 5 and Table 6.

Implementation Detail for reflect pinpoint generation

RG1: MedMCQA Given a multiple-choice question, we employ Prompt I to a LLM, eliciting both an initial response and its reasoning process. The model is required to strictly adhere to a predefined format specification for its final answer. Subsequently, we compare the model's generated final answer option against the standard answer: samples where the options differ are retained, while those that match are discarded. Then, using Prompt II, we identify where in the explanation the erroneous option manifests, isolating the corresponding segment of text containing the critical reasoning flaw. This segment serves as a pinpointed reflection target.

RG2: ChatDoctor Following text preprocessing, we employ Prompt III to guide the model in performing the entity extraction task. Subsequently, we randomly select 1 to 3 recognized entities within a sentence to mask, replacing each with entity type. We then utilize Prompt IV to instruct the model to fill these masked entities. Through multiple retries, we calculate a post-retry similarity score. Entities yielding a score below 0.8 are deemed incorrectly after retries and are retained in their masked state entity type. We repeat this step five times to obtain diverse data.

Implementation Detail for Reflection Generation The generation of reflection questions and answers then proceeds as follows: Generating Reflection Questions: The erroneous masked entity along with the original question are input to the model using Prompt V to generate guiding reflection questions. These questions specifically focus on the relationship between the erroneous entity and the correct answer.

Generating Reflective Answers: Prompt VI is used to direct the model to generate reflective answers based on its own knowledge and capabilities,

addressing the questions generated in the previous step.

Generating Corrected Candidate Answers: Building on the reflection content from steps 1 and 2, Prompt VII guides the model to produce multiple candidate answers incorporating corrected entities.

Each generation step incorporates a retry mechanism to ensure valid output. Finally, the corrected candidate answers undergo strict validation requiring inclusion of all ground-truth correct entities to be accepted.

Data Filtering To guarantee the quality of the reflection data, we first preprocessed the raw datasets to remove samples with short reasoning paths or content irrelevant to medical question-answering. Following this, we conducted a secondary quality assessment to mitigate potential confirmation bias in the model-generated reflections.

This process entailed feeding the generated reflection questions (R_q) and answers (R_a), together with the model's initial erroneous trajectory, back into the model to generate a revised answer. Crucially, the validity of a reflection was determined by verifying the revised answer against the objective ground truth provided by the datasets, ensuring that the correction was not merely endorsed by the model itself but aligned with the gold standard:

(1) For RG1, we utilized the ground-truth option of the multiple-choice question as the standard. A reflection instance was considered valid only if it guided the model to select the correct option matching the dataset label.

(2) For RG2, the objective was to restore masked medical entities. A reflection was deemed valid if the model, guided by the reflection, correctly predicted the original entity present in the raw medical text.

To further ensure the robustness of the reflection data, we repeated this evaluation process 10 times for each instance. We retained a reflection instance only if the model successfully reached the correct ground truth in at least 8 out of these 10 trials. Instances failing to meet this threshold were filtered out to exclude unstable or stochastic reasoning paths. Ultimately, this filtering process yielded a dataset comprising 36,413 medical consultation records and 21,107 multiple-choice questions.

B Dataset Analysis

The key datasets underpinning this study are MedMCQA and ChatDoctor. As shown in the table 4, each sample in the MedMCQA dataset (21,107 samples) contains only a single reflection instance (Reflect avg = 1.0), with reflections primarily occurring at the sentence level. In contrast, the ChatDoctor dataset is significantly larger (36,413 samples) and features richer, more granular reflection content. It encompasses a total of 55,460 reflection instances, averaging 1.52 reflections per sample. This indicates a substantial number of dialogues within ChatDoctor involve diverse reflective processes. Furthermore, ChatDoctor focuses specifically on reflecting medical entities, enabling precise identification of cognitive biases at the entity level and capable of generating 1 to 3 distinct reflection points per response. This structural disparity between the datasets provides the essential data foundation for training models to understand and generate reflections at varying levels of granularity.

C More Implementation Details

To ensure the reproducibility of MedReflect and address the specifics of our training and inference mechanisms, we provide detailed configurations of training.

C.1 Training Details.

Training Data Statistics We tailored the scale of the fine-tuning dataset to the capacity of the target models to ensure optimal adaptation. Specifically, we utilized 2k training examples for the Qwen2.5-7B-Instruct model. For the larger Qwen2.5-32B-Instruct model, we scaled up the training set to 30k examples to prevent under-fitting and ensure robust performance.

Configuration of Supervised Fine-Tuning (SFT). We fine-tuned the base models (Qwen2.5-7B/32B-Instruct) using the generated MedReflect dataset. Table 7 details the specific Configurations used in our experiments. We utilized the AdamW optimizer with a cosine learning rate scheduler. The training was performed on 4 $\times$ NVIDIA A800 (80GB) GPUs.

Metric	MedQA	PubMedQA	MMLU-Pro		GPQA	
			Health	Biology	Genetics	Molecular Biology
Avg. Generated Tokens (All)	711.18	321.31	641.96	721.19	840.35	730.50
Avg. Generated Tokens (Correct)	711.39	320.14	638.30	674.78	904.73	697.90
Avg. Generated Tokens (Incorrect)	710.90	322.20	645.94	810.58	761.67	761.58

Table 3: Statistics of average response length (number of tokens) across different medical benchmarks.

dataset	Num	Total_reflections	Length
MedMCQA	21107	21107	394.56
ChatDoctor	36413	55460	305.05

Table 4: Data Analysis

Configuration	Value
Global Batch Size	4
Learning Rate	$1e-4$
Epochs	3
Max Sequence Length	4096
Optimizer	AdamW
Weight Decay	0.01
Warmup Ratio	0.03
LR Scheduler	Cosine
Gradient Clipping	1.0
LoRA Configuration	
LoRA Rank (r)	16
LoRA Alpha (α)	8
LoRA Dropout	0.05
Target Modules	q, k, v, o, gate, up, down_proj

Table 7: Detailed configuration for SFT

Prompt Template To ensure the model adheres to the desired instruction-following format and medical reasoning style, we employed a specific template during the supervised fine-tuning stage. The input sequence is formatted as follows:

```
<|im_start|>system
You are a medical expert skilled in answer questions while reflecting. Your reflecting goal is to generate a good reflection to assist you to improve your previous answer and continue to answer the question.
<|im_end|>
<|im_start|>user
{Input Question}
<|im_end|>
<|im_start|>assistant
{Target Response}
<|im_end|>
```

Figure 6: The training prompt template used for MedReflect.

C.2 Evaluation Details

Inference Configuration During the inference stage, we utilized the vllm library. We employed **Nucleus Sampling** to ensure the diversity and robustness of the generated responses. The specific generation parameters are listed in Table 8.

Parameter	Value
Temperature	0.7
Top-p	0.9
Repetition Penalty	1.05
Max New Tokens	4096

Table 8: Inference Configurations

C.3 Evaluation Protocol

To mitigate the variance arising from generation randomness and ensure the reliability of our results, we adopted a multi-run evaluation strategy. For standard benchmarks including MedQA, MMLU, PubMedQA, and MedMCQA, we conducted 3 independent runs and reported the average accuracy. For the GPQA dataset, given its relatively small test set size which is susceptible to higher statistical fluctuation, we increased the number of independent runs to 5 and reported the average performance to ensure a more stable assessment.

C.4 Answer Extraction and Verification Strategy

To standardize the evaluation process and mitigate discrepancies arising from formatting inconsistencies in model outputs, we implemented a robust two-stage protocol for answer extraction across all datasets.

Stage 1: Deterministic Pattern Matching. Initially, we employed a strict rule-based extraction mechanism using regular expressions. We defined a set of heuristic patterns to capture conclusive answers, such as matches for `"The answer is option [X]"` or `"My answer is [X]"`. If a distinct option key (A, B, C, or D) was successfully extracted, it was directly compared with the ground truth label.

Stage 2: LLM-Assisted Semantic Parsing. To address instances where the model's response structure was irregular or failed the regex matching (i.e., the answer was implicitly embedded in the reasoning chain without a standard prefix), we utilized a superior-capability LLM (GPT-4) as a semantic parser. This acts as a fallback mechanism to interpret the semantic intent of the model's output.

Specifically, we utilized the `gpt-4-0613` version via API with a temperature setting of 0 to ensure deterministic and reproducible outputs. The model was tasked to act as an objective evaluator. We constructed a specific prompt template (detailed in Table 10) that includes the original question, the candidate options, and the model's raw response. GPT-4 was instructed to identify the final chosen option and output a single character. This hybrid approach ensures that valid responses are not penalized due to minor formatting deviations, thereby providing a more accurate assessment of the model's reasoning capabilities.

C.5 Evaluation setup of Experiment in Section 5.6.2

In the experiments presented in Section 5.6.2, the reflections (comprising reflection questions and answers) within the MedReflect-7B responses were extracted. This reflection information was subsequently utilized as context and concatenated with the original query to directly input into the original Qwen2.5-7B-Instruct. The prompt template used for this inference is shown in Table 9.

Component	Content Details
System Role	You are an excellent medical exam assistant, skilled at extracting correct information from complex responses.
Instruction	I will provide you with a question and the response from a large model. Please help determine which answer the large model believes is correct. If you think it does not answer the multiple-choice question, return None. No explanation is needed, just return the option itself.question: questionanswer: ans
Input Format	**Question:** [Insert Question Text] **Options:** (A) [Option A] (B) [Option B] (C) [Option C] (D) [Option D]

Table 10: The prompt template employed for the GPT-4 based answer extraction stage. The model acts as a semantic parser to extract the final answer key from unstructured reasoning chains.

D More Qualitative Analysis

D.1 Analysis of the Efficiency Boundary for Reflection Data

Dataset	Reflect_2k	Reflect_30k
MedQA	75.5	74.3
PubMedQA	75.3	74.0
MMLU-health	62.8	61.0
MMLU-biology	75.8	72.9
GPQA-Genetics	65.0	63.0
GPQA-Molecular Biology	60.3	59.5

Table 11: Results of the Data Efficiency Experiment of MedReflect-7B

We investigated the impact of training data magnitude on model performance by conducting a comparative analysis between a compact dataset of 2,000 samples and a significantly larger dataset of 30,000 samples. As indicated by the results shown in Table 11, the model trained on only 2,000 samples achieves a performance level comparable to that of the model trained on 30,000 samples. Across various metrics, the smaller dataset yields results commensurate with those of the larger dataset. This observation underscores the exceptional data efficiency of the proposed reflection mechanism, suggesting that the model can effectively internalize the core patterns of reflective reasoning with a limited volume of high-quality instruction data.

D.2 Response Length.

We further report the response length of MedReflect, measured by the number of generated tokens (excluding the prompt). Table 3 summarizes the average length over all samples, as well as the average length of correct and incorrect responses. On MedQA and PubMedQA, correct and incorrect responses have almost the same average length. On MMLUPro-health, correct responses are slightly shorter on average. In contrast, on MMLUPro-biology and GPQA-Molecular Biology, incorrect responses are notably longer than correct ones. On the highly specialized GPQA-Genetics benchmark, correct responses are substantially longer than incorrect ones (904.7 vs. 761.7 tokens on average), suggesting a stronger length–correctness association on this dataset. Overall, response length appears dataset-dependent under MedReflect.

E Case Study

Cases for Multi-Choice Question The multi-choice cases are presented in Figure 7, where we compare the performance of MedReflect, Huatuogpt-o1, and Deepseek-R1.

I We conducted a specific analysis of the performance of MedReflect and the general reasoning large model (DeepSeek-R1), using septicemia pathogen identification as a representative clinical scenario. Experimental results reveal that both DeepSeek-R1 and MedReflect initially exhibit similar diagnostic errors when confronted with complex clinical evidence. However, a critical divergence emerges in their subsequent reasoning pathways:

- DeepSeek-R1, constrained by its reasoning paradigm, terminates analysis upon reaching an incorrect conclusion. It demonstrates no self-correction, maintaining confidence in its initial output.
- MedReflect, leveraging its intrinsic reflective architecture, critically evaluates morphological inconsistencies – questioning the mismatch between observed coccobacilli features and Neisseria gonorrhoeae's typical diplococcal structure. It further assesses clinical epidemiology to validate diagnostic hypotheses.

This structured self-reflection mechanism enables MedReflect to iteratively retrieve evidence from the knowledge base and dynamically adjust diagnostic weights, thereby achieving autonomous error correction and precise diagnosis. In contrast, DeepSeek-R1 despite engaging in constant deep thinking, lack a mechanism for reflection and adjustment of initial responses.

II We conducted a specific analysis of the reasoning differences between MedReflect and Huatuogpt-o1, using mass conversion problems as a representative stoichiometric scenario.

- MedReflect first incorrectly identified ethane. Then, MedReflect exhibited dynamic error correction capability. When significant deviations from given options were detected in its computed results, MedReflect implicitly switched to the correct molecular formula for recalculation, ultimately arriving at the correct answer (G).
- Huatuogpt-o1 manifests a fundamental flaw in its misconception of core principles,in which case it misinterprets the stoichiometric ratio. This foundational error causes the entire computational framework to deviate from theoretical benchmarks. Then, when the calculated result (122.4g) shows significant deviation from options, the model arbitrarily selects the closest match (I) without verifying the validity of stoichiometric assumptions, which indicates a lack of scientific verification or reflection.

Overall, MedReflect tends to follow the physician-like reasoning process and adjust its reasoning process better. In contrast, HuatuoGPT-o1 failed to promptly correct its initial error, compounding the mistake until it ultimately settled for an answer it deemed a close approximation.

Cases for Consulting Question Using a diagnostic case of pediatric respiratory infection presenting with lingual symptoms, we evaluated the performance of three models: DeepSeek, MedReflect, and HuatuoGPT-o1. The performance is shown in Figure 8.

- MedReflect demonstrated more rigorous clinical reasoning through its dynamic reflection mechanism. After generating an initial response, it actively challenged its preliminary assumptions. This process involved critically analyzing relevant pathological mechanisms and prioritizing non-infectious etiologies. Its treatment recommendations appropriately advocated for conservative interventions such as avoiding irritants and dietary modifications.

Figure 7: Cases of Multi-Choice Questions

- HuatuoGPT-o1 also engaged in deep reasoning and reflection during its analysis. However, its reflective process was comparatively superficial. Rather than conducting an in-depth exploration of disease etiology, it employed overly simplistic reasoning, such as excluding diagnoses based solely on "the child not complaining of pain".

- DeepSeek-R1 systematically listed potential diagnoses including lingual papillitis, oral thrush, and viral exanthems. However, its diagnostic suggestions exhibited a scattered approach, failing to effectively integrate core clinical clues. Of course, as a powerful general reasoning model, it may be difficult for it to directly offer suggestions, so it provides many options for patients to choose.

MedReflect's iterative reasoning model, grounded in clinical clue integration, differs significantly from DeepSeek's divergent deep reasoning. It more closely aligns with medical differential diagnostic logic. Simultaneously, its reflective process demonstrates greater rigor and professionalism compared to HuatuoGPT-o1's relatively simplistic and cursory approach.

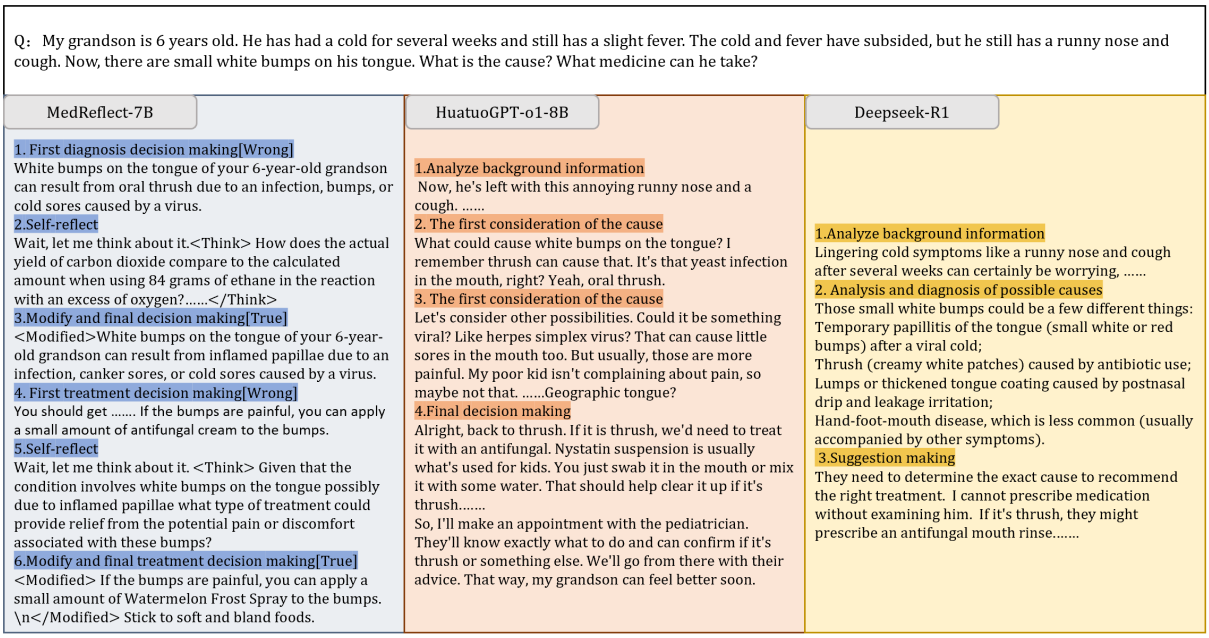

Figure 8: Case of Consulting Question

ID	Detail
Prompt I	You are a medical expert skilled in multiple choice question answering with thorough reasoning. <Query>:[Question]</Query> If you need to specify an option, always use option A/B/C/D. Now, conclude your answer [Therefore, my answer is: option A/B/C/D.] after thorough reasoning:
Prompt II	You are a text processor skill in spliting text according to the option. Split the text into prefix and suffix, the prefix is the content the first time the option is explained in the text. Do not alter the original text. <Text>[Text to be segmented]</Text> <Option>[Incorrect option letter]</Option> <Prefix>
Prompt III	You are a professional doctor. I will provide you with a question and part of the answer. Please read the answer and extract the relevant entities that appear in the answer. Extract professional core terms with high clinical medical characteristics from the perspective of professional doctors, such as etiology, disease, diagnosis, medical examination and testing, medicine, medical therapy and other medical entities. Find all the medical entities that you think suitable in the answer. If you think there are no suitable entities, just ok to return null. Please provide it to me in JSON format, such as [entity type:xxx,entity name:xxx]. No explanation is needed, just give me JSON in English.
Prompt IV	Here is a medical query from your patient: <Query>:[question]</Query> Here is the answer template for the medical query: <Answer template>:[temp text]</Answer template> Please complete the mask section based on the template I provided. The case format for the mask section is as follows: < mask, type: TYPE>, The TYPE is a hint I gave you, implying the entity type in the mask section. Each part of the mask can only be filled with one answer, which can be a word or phrase. You need to find all the mask entities in this sentence and give me the answer you generated, with <mask, type: TYPE, answer: YOURANSWER> is returned without explanation. The TYPE must be the same as the TYPE I gave you. Just give me the return format. Your must give me a answer. And it cannot be the same as the TYPE. Now you can answer here:
Prompt V	Here is a medical query from your patient: <Query>:[query]</Query> Here is your response for the medical query: <Response>:[sentences]</Response> (<correct> [your correct preceding content] </correct>,<incorrect>[your wrong answer] </incorrect>) The initial answer [retry answer] in your response is incorrect, so you need to ask a reflective question based on your correct preceding content (if any) and your wrong answer. Now please provide a brief question(Strictly follow this format:<Reflective Question>your response</Reflective Question>):

Table 5: Prompts for Data Generation (Part 1)

ID	Detail
Prompt VI	Here is a medical query from your patient: <Query>:[query]</Query> Here is your response for the medical query: <Response>:[sentences]</Response> Here is your own reflective question for your response: <Reflective Question>:[Reflective Question]</Reflective Question> (<correct> [your correct preceding content] </correct>,<incorrect>[your wrong answer] </incorrect>) Now, please provide a concise answer for the reflective question(Strictly follow this format:<Reflective Answer>your response</Reflective Answer>):
Prompt VII	Here is a medical query from your patient: <Query>:[query]</Query> Here is your own reflection on your initial answer: <Self-Reflection>:[reflect question][reflect answer]</Self-Reflection> Here is the response you need to complete(Complete each blank): <Response>:[sentences]</Response> Now, according on this reflection, your completed answer is(Strictly follow this format:<Answer>:your refine entity,eg:[weight reduction,cancer]</Answer>):

Table 6: Prompts for Data Generation (Part 2 - Continued)

Prompt	You are an experienced doctor. Please answer the following medical questions. Conclude your response with 'Therefore, my answer is **your option letter**'. For example, if the your answer is option A, you should conclude your response with 'Therefore, my answer is **A**.'. <Query>:[Question]</Query> <Internal Thinking>[reflection_info]< /Internal Thinking> [Warning: The <Internal Thinking> represents your internal thoughts about the <Question>, it's not always correct, it's only for your reference.]

Table 9: Prompts for the experiments in Section 5.6.2. Depending on the experimental setting, `reflection_info` may include both the reflection question and answer, only the reflection question, or only the reflection answer.