EMR: Self-Evolving Medical Multi-Agent System via Experience Mining and Reuse

Dongsheng Shi[1], **Yue Li**[1], **Xin Yi**[1], **Linlin Wang**[1,2*]
[1]East China Normal University
[2]City University of Hong Kong
{dongsheng, yue_li, xinyi}@stu.ecnu.edu.cn,
llwang@cs.ecnu.edu.cn

Abstract

Large language model (LLM) driven multi-agent systems have shown promise in complex clinical reasoning, yet existing approaches rely on static strategies and lack persistent clinical memory, preventing self-evolving from prior diagnostic successes and failures. We present EMR, a self-evolving medical multi-agent system via **E**xperience **M**ining and **R**euse. EMR introduces a hierarchical clinical experience library that organizes accumulated knowledge into three levels: clinical principles, diagnostic patterns, and representative cases. During inference, EMR emulates multidisciplinary consultation: a planner agent coordinates domain-specific department agents for specialized reasoning, while a summary agent synthesizes their analyses into a final decision. Critically, EMR automatically extracts correct diagnostic insights and failure-related warnings from multi-agent reasoning trajectories, incrementally updating the experience library to guide future cases. Experiments on medical reasoning benchmarks demonstrate that EMR consistently outperforms state-of-the-art medical multi-agent baselines. Further analysis reveals that the hierarchical experience enables cross-specialty generalization and transfer across diverse LLM backbones, offering a scalable and interpretable pathway toward continually evolving intelligent medical reasoning.

1 Introduction

Recent advances in large language models (LLMs) have substantially improved their capabilities across a wide range of tasks, particularly in reasoning, planning, and tool use (Yi et al., 2025; Li et al., 2025; Chen et al., 2026; Zhang et al., 2026b). This progress has enabled the rise of autonomous agents and subsequently stimulated the development of multi-agent systems, where multiple agents work together to address complex tasks

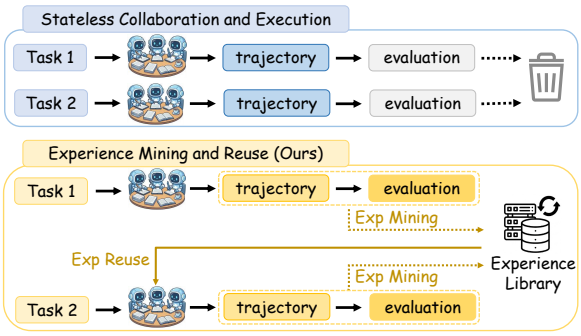


Figure 1: Illustration of the experience mining and reuse mechanism in EMR versus traditional instance-independent reasoning.

(Significant-Gravitas, 2023; Team, 2023; Hong et al., 2024; Chen et al., 2024b; Shi et al., 2026a). In medical settings, such complexity often arises from multi-symptom presentations, differential diagnosis, and the need for cross-department clinical consultation (Lin et al., 2025b; Shi et al., 2026b). In this context, multi-agent systems have been increasingly explored in the medical domain, as they can decompose complex medical problems into coordinated reasoning processes across multiple specialized agents (Tang et al., 2024; Kim et al., 2024).

However, existing medical multi-agent systems suffer from two key limitations. (1) Most current approaches are **static**, relying on fixed reasoning strategies, which often lead to recurring diagnostic errors when facing complex or previously unseen clinical cases (Lin et al., 2025b; Miao et al., 2026; Spieser et al., 2025). (2) As illustrated in Figure 1, medical multi-agent systems typically treat each task as an **isolated** instance, failing to learn from past successful clinical experiences or to avoid previously encountered diagnostic mistakes (Flesch et al., 2018). This lack of experience accumulation fundamentally limits their ability to improve robustness and reliability in medical reasoning.

We argue that the missing component is an ex-

*Corresponding Author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3736–3752
November 6-10, 2026 ©2026 Association for Computational Linguistics

plicit and persistent experience abstraction. Rather than evolving prompts (Yuksekgonul et al., 2024; Zhang et al., 2025c) or workflows (Lin et al., 2025a; Zhang et al., 2025a; Shang et al., 2024) in isolation, multi-agent systems should extract reusable knowledge from their own reasoning trajectories and store it in a form that is transferable, scalable, and model-agnostic. Such experience should capture not only successful reasoning strategies but also systematic failure modes, enabling agents to avoid repeating past mistakes (Cai et al., 2025; Wu et al., 2025). In clinical practice, such experience naturally corresponds to accumulated diagnostic guidelines, common reasoning heuristics, and representative patient cases learned from prior successes and failures.

To this end, we propose EMR, a medical multi-agent system with experience, which enables self-evolving through experience mining and reuse. The multi-agent architecture of EMR is designed to mirror multidisciplinary clinical decision-making workflows. Specifically, a planner agent assigns relevant medical departments, department agents generate domain-specific reasoning trajectories, and a summary agent integrates cross-department analyses to produce final decisions. During the multi-agent collaboration process, EMR automatically mines both golden and warning experiences from agent trajectories and incrementally updates its experience library.

Crucially, EMR organizes experience into a three-level hierarchy—principles, patterns, and cases—corresponding to clinical guidelines, common diagnostic heuristics, and representative patient cases, respectively. This hierarchical design allows high-level medical reasoning strategies to generalize across tasks while preserving fine-grained clinical details when needed. During inference, agents reuse relevant experiences to guide reasoning, enabling the system to continuously improve clinical reasoning performance without modifying model parameters. We evaluate EMR on multiple medical question answering benchmarks. Experimental results show that EMR consistently outperforms static multi-agent baselines, outperforming MDAgents (Kim et al., 2024) by 3.3%, 2.0%, and 1.5% in average accuracy.

The main contributions are summarized as follows:

- We propose EMR, a medical multi-agent system that supports self-evolving through experience mining and reuse within a hierarchical experience library.
- We empirically demonstrate the effectiveness of EMR on medical question answering benchmarks, highlighting the complementary roles of different experience types.
- We show that experiences mined from different LLMs or datasets are transferable and remain human-readable, providing actionable guidance for improving multi-agent reasoning strategies.

2 Related Work

2.1 Multi-Agent Collaboration in Medical

Recent work has explored multi-agent collaboration as a promising paradigm for improving LLM performance in medical tasks. AI Hospital (Fan et al., 2025) and Self-Evolving Multi-Agent Simulations (Almansoori et al., 2025) construct realistic clinical environments where multiple medical agents interact, adapt, and evolve, enabling systematic evaluation of LLMs under complex clinical workflows. For medical reasoning and decision-making, MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024) show that collaborative LLMs with role specialization, adaptive coordination, and consensus mechanisms can significantly enhance single-LLM reasoning and clinical decision quality, while ReConcile (Chen et al., 2024b) further demonstrates the benefits of consensus among diverse LLMs for robust reasoning. Beyond reasoning, ColaCare (Wang et al., 2025) applies LLM-driven multi-agent collaboration to electronic health record modeling, and nomadic multi-agent systems (Dhasarathan et al., 2024) have been used to quantify and protect privacy in dynamic e-healthcare settings. Overall, these studies highlight the effectiveness of multi-agent collaboration in advancing medical reasoning, simulation, and clinical data modeling.

2.2 Experience-Driven Agent Evolution

Recently, experience-driven evolution has emerged as a promising approach for enabling LLM agents to accumulate and reuse past interaction knowledge. ExpeL (Zhao et al., 2024) stores raw trajectories without structured abstraction, while ReasoningBank (Ouyang et al., 2025) generates parallel trajectories under the same model, and G-Memory (Zhang et al., 2026a) encodes experiences

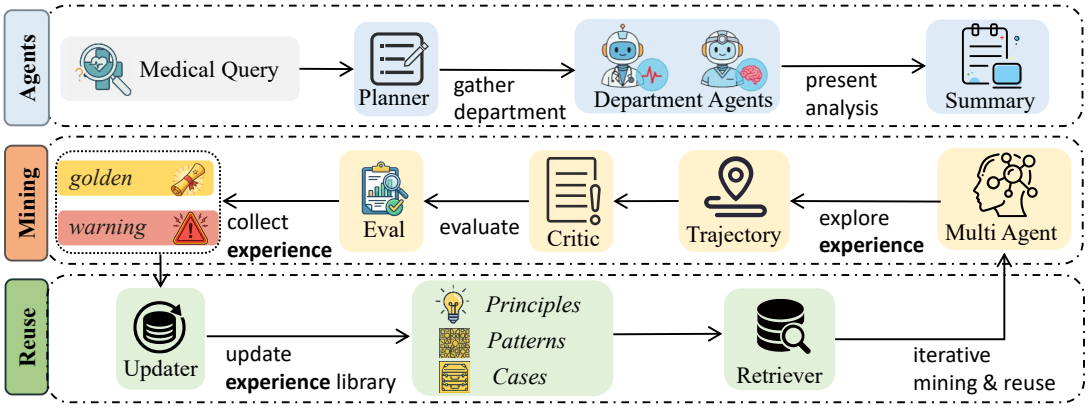


Figure 2: Our proposed EMR diagram. Agents correspond to Section 3.1 and represent the multi-agent collaborative system. Mining corresponds to Section 3.2 and focuses on analyzing multi-agent collaboration trajectories to mine experiential knowledge. Reuse corresponds to Section 3.3, where extracted experiences are integrated into an experience library and reused to guide subsequent tasks, enabling iterative mining and reuse.

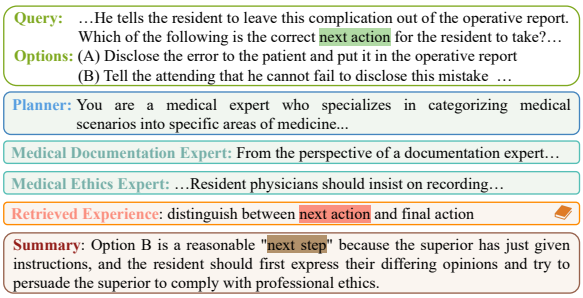


Figure 3: An illustration of EMR reasoning process.

into complex graph structures. In the medical domain, MDTeamGPT (Chen et al., 2025) accumulates consultation experiences as flat summaries for specific agents, whereas EvolveR (Wu et al., 2025) employs GRPO to enhance experience utilization but requires additional computation and parameter updates. Overall, these methods emphasize experience storage or retrieval rather than systematic abstraction and hierarchical reuse within collaborative multi-agent systems, leaving open how accumulated experience can be structured and transferred across diverse agents to support continuous medical reasoning improvement—an issue we address in this work.

3 Method

In this section, we present the components of EMR. As illustrated in Figure 3 and Figure 2, EMR integrates a structured multi-agent collaboration system with an explicit experience library, enabling agents to evolve without updating model parameters. The pipeline comprises three core components: (i) a multi-agent medical system (Section 3.1), (ii) experience mining from collaboration trajectories (Section 3.2), and (iii) experience reuse through hierarchical update and retrieval (Section 3.3).

3.1 Multi-Agent Medical System

Given a medical multiple-choice question q and options $\mathcal{O} = \{o_1, o_2, \ldots, o_K\}$, EMR simulates a multi-department clinical consultation through three types of agents: a planner agent (3.1.1), multiple department agents (3.1.2), and a summary agent (3.1.3).

3.1.1 Planner Agent

Medical questions often span multiple clinical domains, with different options requiring reasoning from distinct diagnostic perspectives (Heist et al., 2014; Stringer et al., 2021). Without explicit domain awareness, downstream agents tend toward redundant or irrelevant exploration (Zhang et al., 2025b; Aryal et al., 2024). To address this, we introduce a planner agent that structures the reasoning space by identifying relevant medical domains at both the question and option levels.

Specifically, the planner assigns question-level domains $\mathcal{D}_q = \{d_1^{(q)}, \ldots, d_M^{(q)}\}$, capturing the core clinical specialties, and option-level domains $\mathcal{D}_o = \{d_1^{(o)}, \ldots, d_N^{(o)}\}$ for evaluating options from diverse perspectives. Incorporating retrieved prior experience $\mathcal{E}_r$ (Section 3.3.3), the planner leverages accumulated patterns to improve domain assignment and avoid previously observed errors. For-

mally:

$$\begin{aligned}\mathcal{D}_q &= \mathrm{LLM}(q, \mathcal{E}_r \mid r_p, \pi_q),\\ \mathcal{D}_o &= \mathrm{LLM}(q, \mathcal{O}, \mathcal{E}_r \mid r_p, \pi_o),\end{aligned} \tag{1}$$

where r_p denotes the planner role, π_q and π_o are prompts for question-level and option-level domain identification, respectively, both incorporating domain taxonomies and task-specific constraints.

By decomposing the reasoning space into domain-specific subproblems, the planner enables downstream department agents to focus on clinically relevant knowledge, laying the groundwork for coordinated collaboration.

3.1.2 Department Agents

While the planner structures the reasoning space, accurate medical decision-making requires in-depth, domain-specific analysis grounded in clinical knowledge and prior experience. We introduce department agents, each emulating a medical expert specialized in a particular clinical domain.

For each domain $d \in \mathcal{D}_q \cup \mathcal{D}_o$, a department agent performs focused reasoning conditioned on the question, candidate options, and retrieved experience:

$$a_d = \mathrm{LLM}(q, \mathcal{O}, d, \mathcal{E}_r \mid r_d, \pi_d), \tag{2}$$

where r_d denotes the domain expert role, and π_d is a domain-specific reasoning prompt guiding the agent to incorporate both medical knowledge and historical experience. All analyses are collected as:

$$\mathcal{A} = \{a_d \mid d \in \mathcal{D}_q \cup \mathcal{D}_o\}. \tag{3}$$

By decomposing medical reasoning into parallel, domain-focused analyses enriched with retrieved experience, department agents provide diverse yet complementary perspectives for the summary agent to integrate.

3.1.3 Summary Agent

While department agents provide parallel, domain-specific analyses, effective medical decision-making requires reconciling potentially complementary or conflicting viewpoints into a coherent global judgment. We introduce a Summary Agent that acts as a coordinating medical consultant, integrating cross-department evidence to derive the final prediction.

Beyond simple aggregation, the summary agent performs reflective reasoning by jointly considering the department analyses and retrieved experience, enabling it to resolve contradictions and correct local errors using accumulated prior experiences. Formally:

$$\hat{y},\ a_{\text{sum}} = \mathrm{LLM}(q, \mathcal{O}, \mathcal{A}, \mathcal{E}_r \mid r_s, \pi_s), \tag{4}$$

where $\hat{y}$ is the predicted answer, a_{sum} the synthesized analysis, r_s the summary role, and π_s an aggregation prompt.

By jointly leveraging multi-department analyses and experience retrieval, the summary agent produces a globally consistent reasoning outcome that reflects both distributed expert knowledge and accumulated experiential insights.

3.2 Experience Mining

EMR enables self-evolving by explicitly mining experiences from multi-agent collaboration trajectories. This process involves trajectory generation (3.2.1) and experience collection (3.2.2).

3.2.1 Multi-Agent Collaboration Trajectory Generation

To support experience-driven evolution, EMR records the entire multi-agent reasoning process as a structured trajectory. In medical settings, learning from experience critically depends on understanding how decisions are made, which intermediate judgments contribute to success or failure, and where errors originate.

For each sample, EMR produces a complete reasoning trajectory:

$$\tau = \big(q, \mathcal{O}, \mathcal{D}_q, \mathcal{D}_o, \mathcal{A}, a_{\text{sum}}, \hat{y}\big), \tag{5}$$

where the planner specifies relevant domains, department agents generate domain-specific analyses, and the summary agent synthesizes cross-department reasoning into a final decision. This trajectory captures intermediate reasoning steps and their organizational structure, forming the fundamental unit for subsequent experience mining.

To distinguish effective from erroneous reasoning, each trajectory is evaluated against the ground-truth answer y. A trajectory is successful if $\hat{y} = y$, and failed otherwise. This binary outcome provides a supervision signal enabling EMR to extract reusable experience from both correct reasoning paths and informative failures.

3.2.2 Experience Collection and Categorization

Reasoning trajectories alone do not directly constitute reusable knowledge. EMR distills trajectories

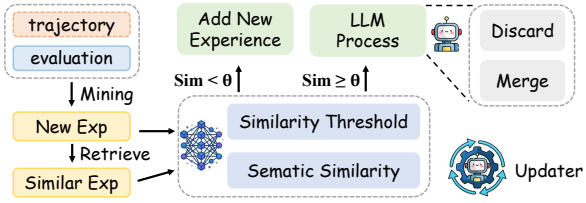


Figure 4: The process of updating the experience library.

into compact, transferable experience representations that abstract away case-specific details while preserving critical reasoning signals. Learning in medical reasoning requires leveraging both successful decision patterns and informative failures that reveal systematic weaknesses.

Given a completed trajectory τ, EMR extracts experience entries ϵ in natural language that encode reusable reasoning knowledge. We categorize experiences into two complementary types:

- **Golden Experience**: distilled from successful trajectories, capturing effective reasoning strategies, diagnostic principles, and validated decision patterns.
- **Warning Experience**: distilled from failed trajectories, summarizing failure causes, misleading cues, and recurring error patterns to avoid in future reasoning.

This dual categorization allows EMR to learn symmetrically from both successes and failures, preventing overfitting to correct outcomes while promoting robust error awareness. Formally:

$$\epsilon = \text{Collect}(\tau, y \mid r_c, \pi_c), \tag{6}$$

where r_c denotes the collector role and π_c the experience mining prompt. The collector compares the trajectory outcome with the ground-truth label y, identifies salient reasoning behaviors or failure modes, and abstracts them into experience entries for long-term accumulation.

3.3 Experience Reuse

EMR supports continual evolution by systematically reusing accumulated experiences during inference, comprising three components: hierarchical experience library (3.3.1), update mechanism (3.3.2), and experience retrieval (3.3.3).

3.3.1 Hierarchical Experience Library

A key challenge in experience-driven multi-agent systems is accumulating knowledge without uncontrolled growth or sacrificed generalization. EMR adopts a hierarchical experience library $\mathcal{L}$ with three abstraction levels:

$$\mathcal{L} = \{\mathcal{L}^{\text{principle}}, \mathcal{L}^{\text{pattern}}, \mathcal{L}^{\text{case}}\}. \tag{7}$$

At the highest level, $\mathcal{L}^{\text{principle}}$ stores abstract diagnostic rules and high-level reasoning principles. The intermediate level $\mathcal{L}^{\text{pattern}}$ captures reusable reasoning structures and decision templates. At the lowest level, $\mathcal{L}^{\text{case}}$ contains concrete, context-specific examples preserving detailed clinical reasoning traces. This hierarchical organization balances abstraction and specificity, enabling both compact storage and flexible reuse.

3.3.2 Experience Update Mechanism

Given a newly generated experience ϵ, the updater agent determines whether and how to incorporate it into the library. The core objective is to preserve coherence and compactness while allowing novel experience to accumulate. Equivalent or redundant entries are discarded or merged rather than blindly added.

The update process consists of three steps. First, the updater retrieves potentially related experiences:

$$\mathcal{R} = \text{Retrieve}(\mathcal{L}_t, \epsilon), \tag{8}$$

where $\mathcal{L}_t$ denotes the library at step t.

Second, the updater estimates semantic relevance between ϵ and $\mathcal{R}$. Each experience is encoded into a semantic vector, and similarity is computed via cosine similarity:

$$\text{Sim}(\epsilon, \mathcal{R}) = \max_{\epsilon_i \in \mathcal{R}} \frac{\mathbf{e}(\epsilon) \cdot \mathbf{e}(\epsilon_i)}{\|\mathbf{e}(\epsilon)\|, \|\mathbf{e}(\epsilon_i)\|}, \tag{9}$$

where $\mathbf{e}(\cdot)$ denotes the embedding function.

Finally, the library is updated according to:

$$\mathcal{L}_{t+1} = \begin{cases} \mathcal{L}_t \cup \{\epsilon\}, \text{ if } \text{Sim}(\epsilon, \mathcal{R}) < \theta, \\ (\mathcal{L}_t \setminus \mathcal{R}) \cup \text{LLM_Process}(\mathcal{R}, \epsilon), \text{otherwise.} \end{cases} \tag{10}$$

When similarity falls below θ, the experience is considered novel and added directly. Otherwise, the LLM determines whether to merge them into a more compact representation or discard ϵ if it adds no reasoning value.

This mechanism enables scalable experience accumulation by combining lightweight similarity-based routing with LLM-based semantic consolidation, ensuring the library evolves in a structured and reusable manner.

3.3.3 Experience Retrieval Mechanism

Accumulated experience is only valuable if effectively reused during future reasoning. Rather than statically encoding experience into model parameters, EMR treats the experience library as an external, model-agnostic knowledge source dynamically queried at inference time.

Given a query context c, the retrieval mechanism selects relevant experiences:

$$\mathcal{E}_r = \text{Retrieve}(\mathcal{L}, c, k), \tag{11}$$

where k denotes the number of retrieved entries. Retrieval proceeds hierarchically: the agent first identifies relevant high-level principles, then retrieves associated reasoning patterns, and finally accesses concrete cases grounding abstract knowledge in specific scenarios.

The retrieved experiences are injected into the planner, department, and summary agents as external guidance, shaping reasoning behavior without modifying model parameters. By conditioning reasoning on accumulated experience, EMR enables agents to benefit from prior successes, avoid previously observed failures, and generalize beyond their original training distribution.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate EMR on a diverse set of medical reasoning benchmarks, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU (Hendrycks et al., 2020), and MMLU-Pro (Wang et al., 2024). These datasets cover a wide range of medical problem types, from clinical decision-making to complex medical reasoning, providing a comprehensive testing platform for evaluating medical multi-agent systems.

Implementation. We instantiate EMR with Qwen3-8B (Team, 2025) and GPT-4o (Hurst et al., 2024) to examine robustness across model scales and architectures. The numbers of domain experts for the question and options are set as: $m = 3, n = 2$. Analogous to training epochs, we set the number of data mining iterations to 5 in our implementation. More implementation details see Appendix A. All reported results are averaged over three independent runs.

Baselines. We compare EMR against both single-LLM and medical multi-agent baselines. We use zero-shot prompting (Kojima et al., 2022) as a single-LLM baseline. Multi-agent baselines include MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024), which represent state-of-the-art static medical multi-agent systems. Additionally, we compare against experience-based reasoning methods including ExpeL (Zhao et al., 2024), ReasoningBank (Ouyang et al., 2025), G-Memory (Zhang et al., 2026a), MDTeamGPT (Chen et al., 2025), and EvolveR (Wu et al., 2025).

4.2 Main Results

Comparison with static multi-agent systems. Table 1 shows that EMR consistently outperforms existing medical multi-agent systems across all LLMs and datasets. Under Qwen3-8B, EMR raises the average accuracy from 77.8 (MDAgents) to 81.1; under GPT-4o, it reaches 93.9. Unlike single-LLM baselines, EMR decomposes complex problems and enables structured collaboration. Compared to prior multi-agent methods, EMR introduces explicit experience mining and reuse, avoiding static role assignment and recurring failures.

Comparison with experience-based methods. Among experience-based approaches, EvolveR achieves 92.7 (MedQA) and 86.5 (MedMCQA) via GRPO optimization, while MDTeamGPT reaches 88.6 on MedMCQA with flat consultation summaries. Others (ExpeL, ReasoningBank, G-Memory) lag due to limited abstraction or adaptation. EMR surpasses all, achieving 96.8 on MedQA (+4.1) and 89.2 on MedMCQA (+0.6), with larger gains on complex USMLE-style questions. On MMLU-Pro (Health & Biology), EMR consistently beats all baselines across LLMs. Notably, EMR also boosts smaller LLMs like Qwen3-8B, closing the gap with larger models, while still enhancing strong LLMs such as GPT-4o.e-driven reasoning complements model capacity rather than replacing it.

5 Ablation Study

5.1 Ablation Study on EMR Components

To assess the contribution of each EMR component, we conduct ablation experiments by progressively removing key modules (Table 2). Removing experience mining and retrieval (w/o Exp) causes a 6.7% drop, reducing EMR to the level of static systems like MedAgents. This shows that multi-agent collaboration alone is insufficient without experience accumulation. Among all components,

Table 1: Main results (Accuracy %) on various medical benchmarks. The datasets include MedQA (USMLE), MedMC. (representing MedMCQA, Indian medical exams), PubMed. (representing PubMedQA, biomedical research questions), and various subsets of MMLU and MMLU-Pro, specifically: Ant. (Anatomy), Clin. (Clinical Knowledge), Coll. (College Medicine), Gen. (Medical Genetics), Hea. (Health), and Bio. (Biology). Bold indicates the best performance within each backbone model group.

Method	MedQA	MedMC.	PubMed.	Ant.	Clin.	Coll.	Gen.	Hea.	Bio.	Avg.
Qwen3-8B										
Single-LLM	70.2	66.5	72.1	68.8	71.4	69.5	73.2	60.1	65.8	68.6
MedAgents	75.4	72.3	77.2	76.5	79.1	75.8	78.4	66.5	73.6	75.0
MDAgents	77.8	75.6	80.5	78.9	81.5	77.2	82.1	69.5	76.8	77.8
ExpeL	76.5	73.8	79.2	77.5	80.1	76.8	80.5	68.2	75.1	76.4
ReasoningBank	77.2	74.5	79.8	78.2	80.8	77.5	81.2	68.8	75.8	77.1
G-Memory	77.8	75.1	80.2	78.8	81.4	77.8	81.8	69.2	76.3	77.6
MDTeamGPT	78.5	76.2	81.0	79.5	82.1	78.5	82.6	70.1	77.0	78.4
EvolveR	79.8	77.0	82.1	80.8	83.5	79.8	83.9	71.2	78.1	79.6
EMR (Ours)	**81.0**	**78.8**	**83.5**	**82.2**	**85.4**	**80.9**	**84.7**	**73.1**	**80.5**	**81.1**
GPT-4o										
Single-LLM	86.5	83.1	82.4	85.9	88.1	87.2	89.5	79.2	83.4	85.0
MedAgents	91.5	87.4	89.2	90.5	94.1	91.8	93.2	83.4	88.9	90.0
MDAgents	93.8	90.5	91.6	92.8	96.2	93.5	95.8	85.6	91.4	92.4
ExpeL	87.4	84.1	83.5	86.8	89.2	88.1	90.3	80.1	84.5	86.0
ReasoningBank	88.6	85.5	84.8	88.1	90.5	89.2	91.6	81.3	85.8	87.2
G-Memory	89.3	84.9	85.2	88.5	90.9	89.6	92.0	81.8	86.2	87.6
MDTeamGPT	90.1	88.6	86.5	89.4	92.1	90.8	93.1	82.9	87.5	88.9
EvolveR	92.7	86.5	88.1	91.2	93.8	92.5	94.9	84.2	89.1	90.3
EMR (Ours)	**96.8**	**89.2**	**93.4**	**94.7**	**97.5**	**95.4**	**98.2**	**86.8**	**93.1**	**93.9**

Table 2: Ablation study on MedQA and MedMC datasets with Qwen3-8B. Exp: Experience, Pla: Planner, Dep: Department Agents.

Method	MedQA Acc. (%)	MedMC Acc. (%)
Single-LLM	70.2	66.5
EMR (Full)	**81.0**	**78.8**
w/o Exp	74.3	71.2
w/o Exp + Pla	72.8	69.5
w/o Exp + Pla + Dep	70.5	67.8
w/o Pla + Dep	78.1	75.4

Table 3: Ablation study for the sub-components of the Experience module on MedQA and MedMC datasets with Qwen3-8B.

Method	MedQA Acc. (%)	MedMC Acc. (%)
EMR (Full)	**81.0**	**78.8**
w/o Cases	79.8	77.4
w/o Patterns	78.5	75.9
w/o Principles	76.4	73.5

experience yields the largest performance gain, underscoring its central role in enabling continuous improvement.

Interestingly, when planner and department agents are removed but experience is retained (w/o Pla + Dep), EMR still maintains strong performance, with only a 2.9% decrease compared to the full model. This suggests that accumulated experience can partially compensate for the absence of explicit role decomposition by directly guiding the summary agent's reasoning.

5.2 Ablation Study on Experience Hierarchy

We analyze the impact of different experience hierarchy levels using the ablation results in Table 3. Removing case-level experiences leads to a modest 1.2% drop, indicating that concrete historical cases offer useful but limited guidance. Ablating pattern-

level experiences causes a larger 2.5% degradation, suggesting that abstracted reasoning patterns are more important for generalization across questions. The most significant decline occurs when principle-level experiences are removed, with a substantial 4.6% drop that brings EMR close to baseline performance. These results confirm that principle-level knowledge serves as the core transferable component for robust reasoning across diverse medical scenarios, while all three levels work complementarily to support overall system performance.

6 Analysis

6.1 Experience Transferability

Table 4 shows that experiences mined by stronger models such as GPT-4o and DeepSeek-V3 (DeepSeek-AI, 2024) yield substantial gains when reused by smaller models (e.g., Qwen3-8B), reflecting their richer medical knowledge, more reliable reasoning patterns, and higher-quality failure summaries. Notably, this transferability is bidirectional: even strong LLMs benefit from reusing experiences mined by smaller models, which contribute diverse diagnostic perspectives, alternative reasoning heuristics, and complementary error patterns.

Table 4: The impact of experience from different models on other models.

	w/o Exp	+ Qwen	+ GPT-4o
MedQA			
Qwen3	74.3	81.0 (+6.7)	83.9 (+9.6)
DeepSeek	86.2	92.5 (+6.3)	94.5 (+8.3)
MedMCQA			
Qwen3	71.3	78.8 (+7.5)	81.2 (+9.9)
DeepSeek	79.1	84.6 (+5.5)	86.1 (+7.0)
PubMedQA			
Qwen3	76.2	83.5 (+7.3)	86.2 (+10.0)
DeepSeek	81.5	86.5 (+5.0)	88.4 (+6.9)

Table 5: Cross-dataset experience analysis.

Source	MedQA	MedMCQA
MedQA	**81.0** (+6.7)	74.8 (+3.5)
MedMCQA	77.5 (+3.2)	**78.8** (+7.5)

Table 5 presents a cross-dataset analysis of experience transferability. While experiences mined from the same dataset yield the strongest improvements on their corresponding benchmarks, experiences extracted from one dataset also consistently improve performance when transferred to different datasets. This indicates that the experience library captures generalizable medical reasoning knowledge rather than dataset-specific heuristics. Notably, the accumulated experiences remain effective across datasets with different question styles and distributions, demonstrating that the experience library serves as a transferable and reusable knowledge resource for medical multi-agent reasoning.

6.2 Experience Analysis

Through qualitative analysis of mined reasoning trajectories, we identify several recurring failure patterns in medical multi-agent reasoning. **(1) Reliance on parametric knowledge rather than clinical guidelines** occurs when agents default to internally memorized facts instead of explicitly reasoning with authoritative diagnostic or treatment standards. **(2) Medical concept confusion** is frequently observed, where closely related conditions, symptoms, or clinical terms are incorrectly conflated, leading to flawed intermediate reasoning. **(3) Superficial causal reasoning** arises when agents focus on surface-level correlations while failing to identify the underlying pathophysiological cause of a condition. Finally, **(4) collective hallucination in multi-agent collaboration** emerges when multiple agents reinforce an incorrect assumption without sufficient cross-verification, amplifying shared errors instead of correcting them. Examples see Appendix E.

7 Conclusion

We presented EMR, a self-evolving medical multi-agent system driven by experience mining and reuse. By organizing golden and warning experiences into principle-, pattern-, and case-level knowledge, EMR progressively improves medical reasoning quality and robustness. Experiments across multiple medical benchmarks demonstrate consistent gains over strong multi-agent baselines, while ablation and transfer studies verify the effectiveness and reusability of the experience library. Our results suggest that explicit experience modeling provides a scalable and interpretable framework for improving medical multi-agent systems and experience-driven collaborative LLM agents.

Limitations

Despite the promising performance of EMR, several limitations remain.

First, EMR currently focuses on text-based medical question answering benchmarks and does not incorporate multimodal clinical information such as medical imaging, laboratory signals, or electronic health records. Real-world clinical decision-making often depends on heterogeneous data sources, and extending experience mining and reuse to multimodal settings remains an important direction for future work.

Second, the quality of the experience library is inherently dependent on the reasoning quality of the underlying LLMs. Although EMR can extract both golden and warning experiences, incorrect or hallucinated reasoning trajectories may still introduce noisy experiences into the library. While the similarity-based update mechanism partially mitigates this issue, more robust experience verification and filtering strategies are still needed.

Third, the current experience retrieval mechanism mainly relies on semantic similarity and hierarchical abstraction, which may not fully capture complex causal relationships between medical cases. As the experience library grows, retrieval efficiency and long-term maintenance may also become increasingly challenging.

Finally, although EMR improves reasoning performance on benchmark datasets, our experiments are conducted in controlled offline evaluation settings rather than real clinical environments. The system has not been validated for real-world medical deployment, and its outputs should not be interpreted as clinical advice. Future work should investigate human-in-the-loop evaluation, clinical safety constraints, and collaboration with healthcare professionals before practical adoption.

References

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. 2025. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*.

Shiva Aryal, Tuyen Do, Bisesh Heyojoo, Sandeep Chataut, Bichar Dip Shrestha Gurung, Venkataramana Gadhamshetty, and Etienne Gnimpieba. 2024. Leveraging multi-ai agents for cross-domain knowledge discovery. *arXiv preprint arXiv:2404.08511*.

Zhicheng Cai, Xinyuan Guo, Yu Pei, Jiangtao Feng, Jinsong Su, Jiangjie Chen, Ya-Qin Zhang, Wei-Ying Ma, Mingxuan Wang, and Hao Zhou. 2025. Flex: Continuous agent evolution via forward learning from experience. *arXiv preprint arXiv:2511.06449*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.

Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. 2025. Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*.

Yanyu Chen, Yue Li, Yongyi Cui, Dongsheng Shi, and Lichang Dai. 2026. Reinforcement learning for large language model selective evidence adoption from contaminated retrieval results. *arXiv preprint arXiv:2607.20090*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Chandramohan Dhasarathan, M Shanmugam, Manish Kumar, Diwakar Tripathi, Shailesh Khapre, and Achyut Shankar. 2024. A nomadic multi-agent based privacy metrics for e-health care: a deep learning approach. *Multimedia Tools and Applications*, 83(3):7249–7272.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Timo Flesch, Jan Balaguer, Ronald Dekker, Hamed Nili, and Christopher Summerfield. 2018. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, 115(44):E10313–E10322.

Brian Sanjay Heist, Jed David Gonzalo, Steven Durning, Dario Torre, and David Michael Elnicki. 2014. Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: a mixed methods study. *Journal of graduate medical education*, 6(4):709–714.

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*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

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*.

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 Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yue Li, Xin Yi, Dongsheng Shi, Gerard De Melo, Xiaoling Wang, and Linlin Wang. 2025. Hierarchical safety realignment: Lightweight restoration of safety in pruned large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7600–7612.

Jiaye Lin, Yifu Guo, Yuzhen Han, Sen Hu, Ziyi Ni, Licheng Wang, Mingguang Chen, Hongzhang Liu, Ronghao Chen, Yangfan He, and 1 others. 2025a. Se-agent: Self-evolution trajectory optimization in multi-step reasoning with llm-based agents. *arXiv preprint arXiv:2508.02085*.

Yanna Lin, Shaojie Xu, Wenshuo Zhang, Yushi Sun, Zixin Chen, Yanjie Zhang, and Rui Sheng. 2025b. A survey of llm-based multi-agent systems in medicine. *Preprints*.

Yongming Miao, Jiaxin Wen, Yuemei Luo, and Jin Li. 2026. Medarc: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. *International Journal of Medical Informatics*, 206:106136.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T Le, Samira Daruki, Xiangru Tang, and 1 others. 2025. Reasoningbank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*.

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.

Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. 2024. Agentsquare: Automatic llm agent search in modular design space. *arXiv preprint arXiv:2410.06153*.

Dongsheng Shi, Yue Li, Xin Yi, Yongyi Cui, Huawei Feng, and Linlin Wang. 2026a. Surgent: A surgical multi-agent assistance system across the perioperative workflow. *arXiv preprint arXiv:2605.29368*.

Dongsheng Shi, Xin Yi, Yue Li, and Linlin Wang. 2026b. Benchmarking large language models for end-to-end clinical support in traditional chinese medicine. *Expert Systems with Applications*, page 132267.

Significant-Gravitas. 2023. Autogpt: Build, deploy, and run ai agents. `https://github.com/Significant-Gravitas/AutoGPT`. GitHub repository.

Jackson Spieser, Ali Balapour, Jarek Meller, Krushna Patra, and Behrouz Shamsaei. 2025. Multi-agent ai systems for biological and clinical data analysis. *Preprints*.

JK Stringer, Sally A Santen, Eun Lee, Meagan Rawls, Jean Bailey, Alicia Richards, Robert A Perera, and Diane Biskobing. 2021. Examining bloom's taxonomy in multiple choice questions: students' approach to questions. *Medical Science Educator*, 31(4):1311–1317.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

XAgent Team. 2023. Xagent: An autonomous agent for complex task solving.

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.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, and 1 others. 2025. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pages 2250–2261.

Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and 1 others. 2025. Evolver: Self-evolving llm agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*.

Xin Yi, Yue Li, Shunfan Zheng, Linlin Wang, Xiaoling Wang, and Liang He. 2025. Unified attacks to large language model watermarks: spoofing and scrubbing in unauthorized knowledge distillation. *Knowledge-Based Systems*, page 114295.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. 2026a. G-memory: Tracing hierarchical memory for multi-agent systems. *Advances in Neural Information Processing Systems*, 38:12988–13018.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. 2025a. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*.

Heng Zhang, Yuling Shi, Xiaodong Gu, Haochen You, Zijian Zhang, Lubin Gan, Yilei Yuan, and Jin Huang. 2025b. D3mas: Decompose, deduce, and distribute for enhanced knowledge sharing in multi-agent systems. *arXiv preprint arXiv:2510.10585*.

Kaituo Zhang, Zhen Xiong, Mingyu Zhong, Zhimeng Jiang, Zhouyuan Yuan, Zhecheng Li, and Ying Lin. 2026b. Are tools all we need? unveiling the tool-use tax in llm agents. *arXiv preprint arXiv:2605.00136*.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, and 1 others. 2025c. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

A Implementation Details

For MedQA and MedMCQA, we split the datasets evenly, using half for experience mining and the other half for evaluation. For PubMedQA, 500 instances are used for experience mining and another 500 (official test set) for evaluation. MMLU (anatomy, clinical knowledge, college medicine, medical genetics) and MMLU-Pro (health, biology) are evenly split between experience mining and testing, with 200 instances per selected MMLU-Pro subset. We adopt bge-m3 (Chen et al., 2024a) as the semantic embedding model and set the similarity threshold to 0.85 (See Appendix B for more details). The number of retrieved experience is 3. Statistically, the cost of experience mining is approximately $5 per 100 samples (5 epochs), with a total runtime of about 2 hours, while inference takes around 45 seconds per sample.

B Sensitivity Analysis of Similarity Threshold θ

The similarity threshold θ controls how strictly the retriever filters historical experiences during the update phase. A low threshold admits noisy or irrelevant experiences, while an excessively high threshold may leave the experience bank too sparse to provide meaningful guidance. To determine the optimal operating point, we sweep $\theta \in \{0.70, 0.80, 0.85, 0.90\}$ on the MedQA and MedMCQA validation sets using the Qwen3-8B backbone.

Table 6: Sensitivity of the experience update mechanism to the similarity threshold θ and ablation on experience types (Qwen3-8B).

Setting	MedQA	MedMCQA
$\theta = 0.70$	80.1	77.9
$\theta = 0.80$	80.7	78.5
$\theta = 0.85$ (default)	81.0	78.8
$\theta = 0.90$	80.5	78.3
Golden-only	78.4	76.2
Golden + Warning (Ours)	81.0	78.8

Threshold sensitivity. As shown in Table 6, $\theta = 0.85$ yields the best performance on both benchmarks (81.0 on MedQA and 78.8 on MedMCQA). When θ drops to 0.70, accuracy falls by 0.9 and 0.9 points, respectively, indicating that overly permissive retrieval introduces distractive or low-quality experiences that mislead the reasoning process. Conversely, raising θ to 0.90 degrades performance by 0.5 and 0.5 points, suggesting that excessive filtering prunes useful cases and reduces coverage of the experience bank. The sweet spot at 0.85 strikes a balance between precision and recall

of retrieved experiences.

Ablation on experience types. We further ablate the contribution of warning experiences. Using only golden (correct) experiences results in a substantial drop of 2.6 points on MedQA and 2.6 points on MedMCQA compared to our full setting. This confirms that warning experiences—capturing failed reasoning patterns and error-prone knowledge boundaries—are complementary to golden ones. They act as guardrails that help the model recognize and avoid previously encountered mistakes, thereby improving robustness on challenging medical questions.

C Prompts

C.1 Prompt for Medical Experience Collection from Multi-Agent Reasoning

You are a senior medical education expert specializing in mining reusable medical experience from multi-agent clinical reasoning trajectories.

Task Description
Given a complete multi-agent medical reasoning trajectory, you are required to:
- Determine whether the final prediction is correct.
- Analyze the reasoning behaviors throughout the multi-agent collaboration process.
- Extract structured and reusable medical experience.

Reasoning Trajectory
{trajectory_text}

Evaluation Criteria
- Ground-truth answer: {golden_answer_idx} ({golden_answer})
- Model prediction: {pred_answer}
- Trajectory status: {"Successful" if is_success else "Failed"}

Experience Collection Requirements
Based on the above trajectory, please extract the following type.

Experience Levels
Principle
High-level diagnostic rules and clinical guidelines. Stable, abstract, and broadly applicable. Limited to 30–50 words.
Pattern
Reusable reasoning templates and decision-making structures describing how to think and analyze. Limited to 30–50 words.
Case
Concrete clinical scenario examples. {"Successful" if is_success else "Failure"}-oriented analysis. Limited to 40–60 words.

Important Notes
- Analyze the complete reasoning trajectory, including Planner, Department Agents, and Summary Agent.
- Pay attention to consensus formation and disagreement resolution.
- Extract abstract and reusable experience rather than restating the original question.

Output Format
{
"is_correct": "{str(is_success).lower()}",
"principle": "...",
"pattern": "...",
"case": "..."
}

C.2 Prompt for Planner Agents

You are a medical expert who specializes in categorizing medical scenarios into specific areas of medicine. You can leverage accumulated reasoning patterns from past experiences to improve domain assignment and avoid previously observed errors. You need to complete the following steps:

1. **Carefully read the medical scenario presented in the question:**

'''{question}'''
Retrieved Relevant Experience
The following experiences are retrieved from the experience library. Please consider them when assigning domains:
Principles
{principles_str}
Patterns
{patterns_str}
Cases
{cases_str}

Note: If the above experiences are not relevant to the current question, please ignore them; if they are valuable, please incorporate them into your domain assignment.

2. **Based on the medical scenario, classify the question into {num_domains} different subfields of medicine.**
3. **You should output in exactly the same format as:**

'''{domain_format}'''

Output only the formatted result, no explanation needed.

C.3 Prompt for Department Agent Medical Analysis

You are a senior {department} expert with extensive clinical experience and profound medical knowledge. Based on your professional background and the retrieved relevant experiences, please conduct an in-depth analysis of the following medical multiple-choice question.
Question Information
Question:
{question}
Options:
{options}
Retrieved Relevant Experiences
The following experiences, including successful strategies (Golden) and lessons from failures (Warning), have been retrieved from the experience library. Please incorporate them into your analysis:
{principles_str}
{patterns_str}

{cases_str}
Note:
- **Golden** experiences demonstrate correct reasoning paths and verification strategies.
- **Warning** experiences highlight common error patterns and pitfalls to avoid.
- Combine both perspectives: leverage successful methods while proactively avoiding known errors.

Please use professional and rigorous medical language, demonstrating both your expertise and clinical reasoning process.

C.4 Prompt for Summary Agent

You are a senior medical diagnostic expert specializing in integrating multidisciplinary opinions to make a final judgment. Based on the analysis from various department experts and the retrieved relevant experiences, please provide a comprehensive assessment for this medical multiple-choice question.
Question Information
Question:
{question}
Options:
{options_str}
Departmental Expert Analyses
{all_analysis}
Retrieved Relevant Experiences
The following experiences, including successful strategies (Golden) and lessons from failures (Warning), have been retrieved from the experience library. Please refer to these during your synthesis:
{principles_str}
{patterns_str}
{cases_str}
Note:
- **Golden** experiences demonstrate correct reasoning paths and verification strategies.
- **Warning** experiences highlight common error patterns and pitfalls to avoid.
- Leverage successful methods while proactively avoiding known errors mentioned in Warning experiences.
Comprehensive Analysis Task
Please complete the following steps:

1. **Synthesized Judgment:** Weigh departmental opinions according to their professional relevance. Identify diagnostic traps or common errors (referencing Warning experiences) and apply the correct reasoning patterns from Golden experiences.
2. **Final Answer:** Clearly state the correct option (must be one of {list(options.keys())}), summarize the core rationale for your choice, and explain how to prevent common mistakes related to this problem.

Output Format
Please output in the following format:
[Comprehensive Analysis]
(Provide a detailed comprehensive analysis, including opinion consolidation, evidence analysis, and the judgment process.)
[Final Answer]
(Output only the option letter, e.g., A)

D Case Study

Question:
A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take?

(A) Disclose the error to the patient and put it in the operative report
(B) Tell the attending that he cannot fail to disclose this mistake
(C) Report the physician to the ethics committee
(D) Refuse to dictate the operative report

Experts: Medical Ethics, Orthopaedic Surgery, Medical Documentation

Analyses from various experts:

Medical Ethics Expert: Resident physicians should insist on recording complications in the surgical record and disclosing them to the patient. This is a direct reflection of upholding patient autonomy, ensuring medical transparency, and fulfilling professional responsibility. If superiors object, resident physicians should still insist and, if necessary, seek support from the hospital ethics committee.

Orthopaedic Surgery Expert: As a senior orthopedic specialist, I emphasize that the growth of resident physicians includes learning to uphold professional standards under pressure. In this case, despite pressure from superiors, the correct course of action was to honestly record and inform the patient. This protects both the patient and the physician's own professional integrity.

Medical Documentation Expert: From the perspective of a seasoned medical documentation expert, the completeness of medical records is non-negotiable; any intraoperative event must be recorded truthfully. Furthermore, honesty with the patient is central to the doctor-patient relationship. Faced with an erroneous instruction from superiors, the correct action is to uphold professional ethics.

Summary:
«Experience Retrieved: When ethical obligations conflict with directives from superiors, one must distinguish between 'process steps' and 'final actions'.»
... Option B is a reasonable "next step" because the superior has just given instructions, and the resident should first express their differing opinions and try to persuade the superior to comply with professional ethics. If the superior still insists, the resident should record and disclose (A), but that is a second step.

E Experience Examples

Below is a detailed example of the experience.

- **Parametric Over-reliance.** An agent recommends immediate PCI for a STEMI patient based on the general belief that "PCI is superior," while neglecting a critical guideline constraint: when the expected delay to PCI

exceeds 120 minutes, fibrinolysis should be the preferred reperfusion strategy.

- **Medical Concept Confusion.** In a legal–medical reasoning task, agents conflate the regulatory definitions of *hospital infection outbreak* and *hospital infection prevalence*, resulting in an incorrect judgment of the mandatory reporting timeframe (12 hours vs. 24 hours).

- **Superficial Causal Reasoning.** When confronted with a limb trauma case presenting numbness and coolness, agents focus on surface correlations and diagnose a nerve injury, failing to identify the underlying pathophysiological cause: acute arterial occlusion, in which neurological symptoms are secondary to ischemia.

- **Collective Hallucination.** During a pharmacology discussion, multiple agents correctly recognize that 5-FU inhibits thymidylate synthase, but reinforce each other's false assumption that its structure mimics thymine (T), collectively overlooking the fact that 5-FU is a uracil (U) analogue.

F Ablation Study on the Number of Agents

This study investigates how the number of agents affects the performance and experience accumulation of EMR. The motivation is twofold: first, to understand whether increasing the number of agents leads to better medical reasoning performance; and second, to examine how agent multiplicity influences the diversity and scale of mined experience. Since EMR relies on collaborative reasoning and experience mining from agent trajectories, the number of agents directly controls the breadth of analytical perspectives available during inference.

As shown in Figure 5, increasing the number of agents generally improves accuracy across all benchmarks, particularly when moving from a small number of agents to a moderate scale. This trend suggests that additional agents introduce complementary medical perspectives, enabling more comprehensive exploration of diagnostic hypotheses and reducing reasoning blind spots. However, performance gains gradually saturate as the number of agents continues to grow, indicating diminishing returns once sufficient analytical coverage is achieved. Beyond performance, increasing the number of agents substantially enlarges the experience library by enriching the diversity of reasoning trajectories, particularly when scaling from a small number of agents.

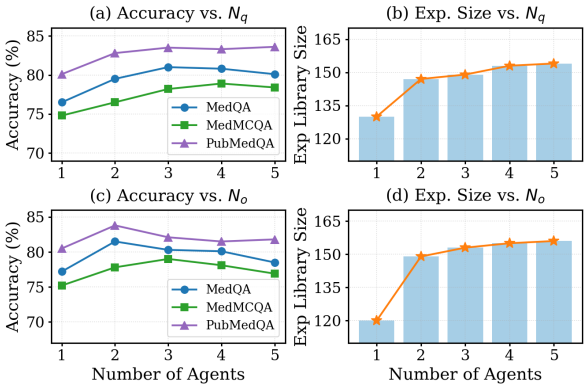


Figure 5: The impact of the number of agents on accuracy and experience library size. The ablation experiments are conducted on the Qwen3-8B model and the MedQA dataset.

G Ablation Study on Retrieved Experience Size

As shown in Figure 6, increasing the number of retrieved experience blocks improves performance in the early stage, with the most significant gains observed when N increases from 1 to 3. This suggests that a small set of diverse and highly relevant experiences is sufficient to effectively guide multi-agent reasoning by reusing successful diagnostic strategies and avoiding previously identified failure patterns.

Notably, the diminishing performance gains beyond this range indicate that EMR benefits more from the quality and abstraction level of reused experience than from sheer quantity. Since experiences in EMR are distilled into reusable principles and patterns, a limited number of experience blocks can already provide strong guidance, reflecting an efficient form of experience reuse rather than simple memory retrieval. From a clinical perspective, this trend aligns with human medical reasoning, where a few representative guidelines or prior cases often suffice to inform diagnosis, while excessive prior information may introduce redundancy or distraction. These findings further suggest that EMR achieves self-evolving not by increasing inference-time context indiscriminately, but by progressively mining and reusing compact, high-level experience that remains effective even under limited retrieval budgets.

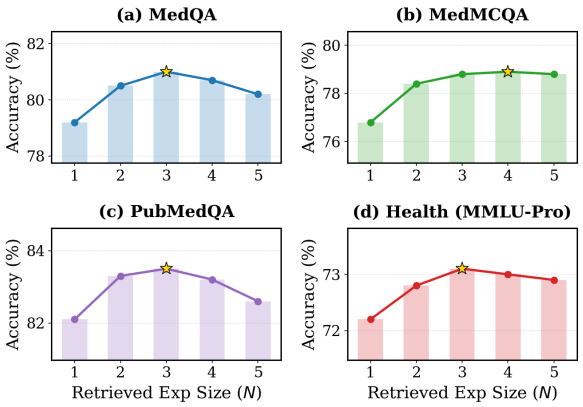


Figure 6: The impact of the number of retrieved experiences on performance. The ablation experiments are conducted using the Qwen3-8B model on the MedQA, MedMCQA, PubMedQA, and Health (MMLU-Pro) datasets.

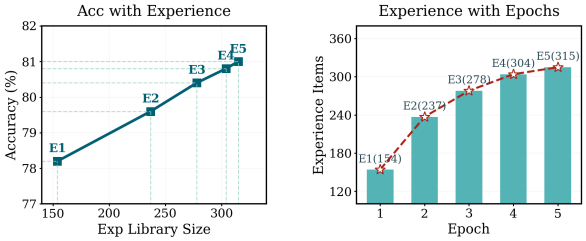


Figure 7: The size of the experience library varies with the number of mining epochs. The analysis experiments are conducted on the Qwen3-8B model and the MedQA dataset.

H Effect of Experience Library Size

We evaluate EMR over multiple experience accumulation epochs, where newly collected experiences from agent reasoning trajectories are incrementally merged into the experience library. We track both the growth of the experience library and the corresponding test accuracy to examine how medical reasoning performance evolves with accumulated clinical experience.

As shown in Figure 7, test accuracy consistently improves as the experience library expands across epochs. Performance gains are particularly pronounced in the early epochs, indicating that newly mined experiences effectively capture high-value clinical knowledge, such as common diagnostic reasoning patterns and frequently observed failure modes, which can be immediately reused to guide subsequent medical decision making. As the number of epochs increases, the discovery rate of novel and informative clinical experiences gradually decreases, leading to a slower growth of the experience library. Correspondingly, performance improvements become more moderate and eventually stabilize. This saturation effect suggests that EMR progressively distills the most salient medical experience embedded in agent trajectories, after which additional accumulation primarily yields redundant or marginal clinical guidance.

I Computational Cost & Efficiency Analysis

As shown in Table 7, both the token consumption and inference latency of EMR remain fully practical, while consistently outperforming existing medical multi-agent baselines in efficiency. Our multi-agent architecture is intentionally designed to be lightweight and efficient. Although inspired by MedAgents, EMR removes the expensive multi-round discussion mechanism commonly used for consensus-reaching, and instead introduces an experience-based conflict resolution strategy. During the entire collaborative reasoning process, the experience module consumes only around 300 additional tokens and requires approximately 2 seconds for retrieval.

In contrast, RECONCILE and MDAgents rely on more complex collaborative pipelines involving iterative reporting or multi-round discussions, which inevitably result in substantially higher token usage and longer inference time. These results demonstrate that explicit experience reuse can improve reasoning quality without introducing excessive computational overhead.

J Generalizability of EMR

To further evaluate the generalizability of the proposed experience mechanism, we integrate EMR into existing medical multi-agent frameworks. To our knowledge, this is the first attempt to incorporate an explicit experience mining and reuse mechanism into medical multi-agent systems. Importantly, EMR is not restricted to a specific architecture and can be naturally extended to any multi-agent framework capable of collecting reasoning trajectories.

As shown in Table 8, integrating EMR consistently improves the performance of both MDAgents and RECONCILE on MedQA and MedMCQA under the GPT-4o backbone. Specifically, MDAgents improves from 93.8 to 96.5 on MedQA and from 90.5 to 92.7 on MedMCQA after incorporating EMR. Similarly, RECONCILE achieves

Table 7: Efficiency comparison of different methods.

Method	Avg. Tokens / Sample	Avg. Time / Sample (s)
Single-LLM (CoT)	200	5
MedAgents (Tang et al., 2024)	1500	18
RECONCILE (Chen et al., 2024b)	2500	33
MDAgents (Kim et al., 2024)	1900	23
EMR (Ours)	**1200**	**15**

gains of 2.7 and 2.4 points, respectively.

These results suggest that the benefits of experience mining and reuse are architecture-agnostic and can serve as a general enhancement strategy for collaborative medical reasoning systems.

Table 8: Generalizability of EMR under the GPT-4o backbone.

Method	MedQA	MedMCQA
MDAgents (Kim et al., 2024)	93.8	90.5
MDAgents (w/ EMR)	**96.5**	**92.7**
RECONCILE (Chen et al., 2024b)	92.1	88.2
RECONCILE (w/ EMR)	**94.8**	**90.6**

K Comparison with Prior Experience-Based Methods

As shown in Table 9, the hierarchical experience library in EMR differs fundamentally from prior experience-based methods in both representation and retrieval strategy, enabling a unified framework that balances generalization and specificity.

Existing approaches typically rely on storing raw reasoning trajectories or flat experience summaries, which often suffer from limited abstraction ability and weak transferability across tasks. In contrast, EMR organizes experience hierarchically into principles, patterns, and cases. Principle-level experiences capture abstract and transferable diagnostic knowledge, improving generalization across diverse medical reasoning scenarios. Pattern-level experiences encode reusable reasoning structures, while case-level experiences preserve detailed clinical contexts that directly support fine-grained reasoning.

Furthermore, prior methods generally retrieve only a few similar instances in a flat manner. EMR instead adopts hierarchical retrieval, progressively retrieving high-level principles, reasoning patterns, and representative cases, thereby enabling both broad reasoning guidance and detailed clinical grounding during inference.

L Sensitivity to Experience Library Size and Quality

As shown in Table 10, EMR remains robust under both limited and noisy experience settings, consistently outperforming standard multi-agent systems even when the experience library is significantly constrained or partially corrupted.

When only 20% of the original experience library is retained, EMR still achieves 76.8 on MedQA and 75.4 on MedMCQA, substantially exceeding the performance of the system without experience. Even under the extreme setting where only 10% of experiences are available, EMR preserves strong reasoning capability, indicating that a small number of high-level principles and reusable reasoning patterns already provide substantial guidance for medical decision-making.

We further evaluate robustness under noisy experience conditions by injecting irrelevant or low-quality experiences into the library. Although performance gradually declines as noise increases, EMR remains relatively stable, achieving 79.3 under 20% noise and 78.4 under 40% noise on MedQA. This robustness mainly stems from the hierarchical organization and experience merging mechanism, which mitigate the influence of noisy or redundant experiences through semantic consolidation and abstraction.

These findings suggest that EMR does not rely solely on large-scale memory accumulation. Instead, the principle-level abstraction and hierarchical refinement mechanisms enable the system to preserve strong reasoning performance even under sparse or imperfect experience conditions.

M Potential Risks

Although EMR is designed as a research framework for improving medical multi-agent reasoning, several potential risks should be acknowledged.

Table 9: Comparison with prior experience-based methods.

Aspect	Prior Methods	EMR (Ours)
Representation	Raw trajectories / flat summaries	Hierarchical (principles, patterns, cases)
Generalization	Limited	Strong (via principles)
Detail Learning	Indirect	Direct (via cases)
Retrieval	Few instances	Hierarchical retrieval

Table 10: Sensitivity to experience library size and quality using the Qwen3-8B backbone.

Setting	MedQA	MedMCQA
Full EMR	**81.0**	**78.8**
Limited Exp (20%)	76.8	75.4
Limited Exp (10%)	76.3	74.7
Noisy Exp (+20% noise)	79.3	76.9
Noisy Exp (+40% noise)	78.4	75.0
w/o Experience	75.3	72.1

First, EMR may generate incorrect or misleading medical reasoning due to hallucinations, incomplete knowledge, or flawed collaboration among agents. Since the experience library accumulates knowledge from model-generated trajectories, erroneous reasoning patterns may propagate across future inferences if not properly filtered.

Second, the reuse of historical experiences may introduce bias amplification. Experiences mined from benchmark datasets or particular LLMs may overrepresent certain medical reasoning styles, disease distributions, or clinical assumptions, potentially reducing robustness when applied to diverse real-world populations or uncommon clinical cases.

Third, EMR may increase users' overreliance on AI-generated medical suggestions. Although the system improves reasoning consistency, it is not a substitute for licensed medical professionals. Incorrect outputs in high-stakes clinical settings could lead to harmful medical decisions if used without expert supervision.

Fourth, experience storage and retrieval mechanisms may raise privacy and security concerns in practical deployments involving real patient data. While our experiments only use publicly available benchmark datasets, extending EMR to real-world clinical environments would require strict compliance with medical data protection regulations and careful anonymization of stored experiences.

Finally, the multi-agent collaboration process substantially increases computational cost and inference latency compared to single-agent systems, which may limit scalability and accessibility in resource-constrained healthcare environments.