TAGS: A Test-Time Generalist–Specialist Framework with Retrieval-Augmented Reasoning and Verification

Anonymous ACL submission

Abstract

Recent advances such as Chain-of-Thought prompting have significantly improved large language models (LLMs) in zero-shot medical reasoning. However, prompting-based methods often remain shallow and unstable, while fine-tuned medical LLMs suffer from poor generalization under distribution shifts and limited adaptability to unseen clinical scenarios. To address these limitations, we present **TAGS**, a test-time framework that combines a broadly capable generalist with a domain-specific specialist to offer complementary perspectives without any model fine-tuning or parameter updates. To support this generalist–specialist reasoning process, we introduce two auxiliary modules: a hierarchical retrieval mechanism that provides multi-scale exemplars by selecting examples based on both semantic and rationale-level similarity, and a reliability scorer that evaluates reasoning consistency to guide final answer aggregation. TAGS achieves strong performance across nine MedQA benchmarks, boosting GPT-4o accuracy by 13.8%, DeepSeek-R1 by 16.8%, and improving a vanilla 7B model from 14.1% to 23.9%. These results surpass several fine-tuned medical LLMs, without any parameter updates.

1 Introduction

Large Language Models (LLMs) have recently demonstrated promising capabilities in medical question answering (MedQA), achieving strong performance on a range of benchmark datasets (Singhal et al., 2025; Jin et al., 2022; Chen et al., 2023a,b; Zhou et al., 2023; Gao et al., 2024). However, despite these advances, recent studies have shown that even state-of-the-art models frequently fail on complex cases requiring deep domain knowledge, multi-step reasoning, and generalization to out-of-distribution clinical scenarios (Xu et al., 2024; Fan et al., 2025; Shi et al., 2024).

To improve LLM reasoning in MedQA, two major research directions have emerged. The first involves prompting-based strategies, such as Chain-of-Thought (CoT) (Wei et al., 2022) and Multi-Agent Systems (MAS) (Chen et al., 2025b,a), which guide LLMs through structured multi-step reasoning or simulate expert collaboration. However, empirical studies (Tang et al., 2025) have revealed that interactive multi-agent reasoning is frequently brittle, with redundant outputs, unstable communication, and poor coordination undermining reliability. Recent benchmarks report high failure rates for such systems, with multi-agent discussions yielding limited improvements over single-agent baselines (Cemri et al., 2025). The second direction focuses on fine-tuning LLMs on domain-specific medical data, yielding specialist models like HuatuoGPT (Chen et al., 2024c) and MedLLaMA (Qiu et al., 2024). Although these models perform well on in-distribution tasks, they tend to overfit their training domains and exhibit limited adaptability to emerging knowledge or unseen questions (Yang et al., 2024b; Ye et al., 2024).

While prior work has made progress on reasoning, retrieval, and domain adaptation, these components are typically developed in isolation and lack integration into a unified, inference-time framework. As a result, existing approaches fall short in three critical aspects for robust MedQA: acquiring up-to-date medical knowledge, supporting diverse and complementary reasoning strategies, and ensuring answer reliability under distribution shift, all without relying on any parameter updates. This motivates our work: designing a structured, test-time reasoning framework that unifies generalist and specialist perspectives with retrieval-augmented prompting and verification.

Our motivation stems from the observation that single-agent prompting often lacks the depth and diversity required for complex medical reasoning, while multi-agent systems tend to produce redundant or inconsistent outputs due to unclear role assignments and weak complementarity. To address these limitations, we propose **TAGS** (Test-time Generalist–Specialist framework with retrieval-augmented reasoning and verification), a modular, inference-only framework that enhances medical question answering through structured reasoning collaboration. At its core, TAGS relies on a Generalist–Specialist Reasoning Collaboration (**GSRC**) module, which facilitates complementary reasoning between a generalist agent and a specialist agent. Each agent brings a different perspective to the problem: the generalist offers broad clinical insights, while the specialist contributes precise, domain-specific reasoning. This collaboration is designed to produce diverse and accurate reasoning paths for challenging medical questions.

To support GSRC, we introduce two auxiliary modules. First, to promote diverse and contextually grounded reasoning, we propose a Hierarchical Retrieval Augmentation (**HRA**) mechanism that leverages external, high-quality medical Chain-of-Thought (CoT) exemplars. HRA operates in two stages: it first retrieves semantically relevant CoT exemplars based on the input question–answer (QA) pair to serve as initial references for both agents. Then, after the agents generate their initial reasoning steps, HRA performs a second round of retrieval based on the generated rationales to obtain reasoning-level exemplars that further guide and enrich subsequent inference. Second, to ensure the reliability of the generated answers along with their reasoning chains, we introduce an Uncertainty-Aware Answer Aggregation (**UAAA**) module. UAAA evaluates the consistency between the reasoning path and the final answer produced by each agent, and aggregates the outputs based on a confidence score and inter-agent agreement, yielding a robust and interpretable final prediction. This design effectively mitigates issues common in previous methods, such as insufficient reasoning depth and diversity in single-agent prompting and redundant or conflicting outputs arising from ambiguous roles in multi-agent systems, ultimately producing more accurate and robust medical reasoning outcomes.

Extensive experiments across nine MedQA benchmarks validate the robustness and adaptability of TAGS under both non-fine-tuned and fine-tuned settings. Our method consistently outperforms strong prompting and agent-based baselines across multiple foundation models, including GPT-4o (18.0% to 31.8%), DeepSeek-R1 (27.2% to 44.0%), and Qwen-2.5-7B (14.1% to 23.9%). Notably, we show that even without retrieving semantically similar exemplars, TAGS maintains strong performance by exposing models to diverse reasoning patterns rather than specific answers.

2 Related Work

2.1 Medical Question Answering

Medical question answering (MedQA) aims to predict accurate answers to domain-specific clinical or biomedical questions, often posed in a multiple-choice format. Existing benchmarks span a variety of formats and reasoning challenges, including clinical exam-style questions, evidence-based inference, and multi-subject distractor-rich scenarios (Pal et al., 2022). Recent approaches leverage large language models (LLMs) or chain-of-thought prompting to enhance reasoning (Singhal et al., 2025). MedCoT (Liu et al., 2024) explicitly integrates multi-step rationale generation with hierarchical expert feedback. Concurrently, biomedical LLMs such as MedLLaMA [1], HuatuoGPT (Chen et al., 2024c), and OpenBioLLM (Pal and Sankarasubbu, 2024) have achieved strong zero-shot or few-shot performance on MedQA benchmarks. However, these models typically rely on direct answer generation and lack explicit mechanisms for multi-agent reasoning or consistency verification. In contrast, our method introduces a retrieval-augmented multi-agent framework that performs staged reasoning and employs a dedicated verifier to assess the reliability of generated answers, promoting robustness and interpretability.

2.2 Retrieval-Augmented Reasoning

Retrieval-augmented reasoning enhances prediction quality, factual consistency, and interpretability by incorporating external knowledge into the reasoning process. Early works such as RAG (Lewis et al., 2020) retrieve text passages to guide open-domain generation, while later methods extend retrieval to more structured forms, such as few-shot demonstrations (Izacard et al., 2023) or intermediate reasoning paths (Shi et al., 2023). In the context of chain-of-thought (CoT) prompting, retrieval has been explored to select relevant questions, rationales, or multi-step reasoning exemplars that better align with the target task (Xu et al., 2022; He et al., 2025). Despite these advances, most existing frameworks rely primarily on question-level similarity and often neglect deeper alignment

[1] https://huggingface.co/johnsnowlabs/

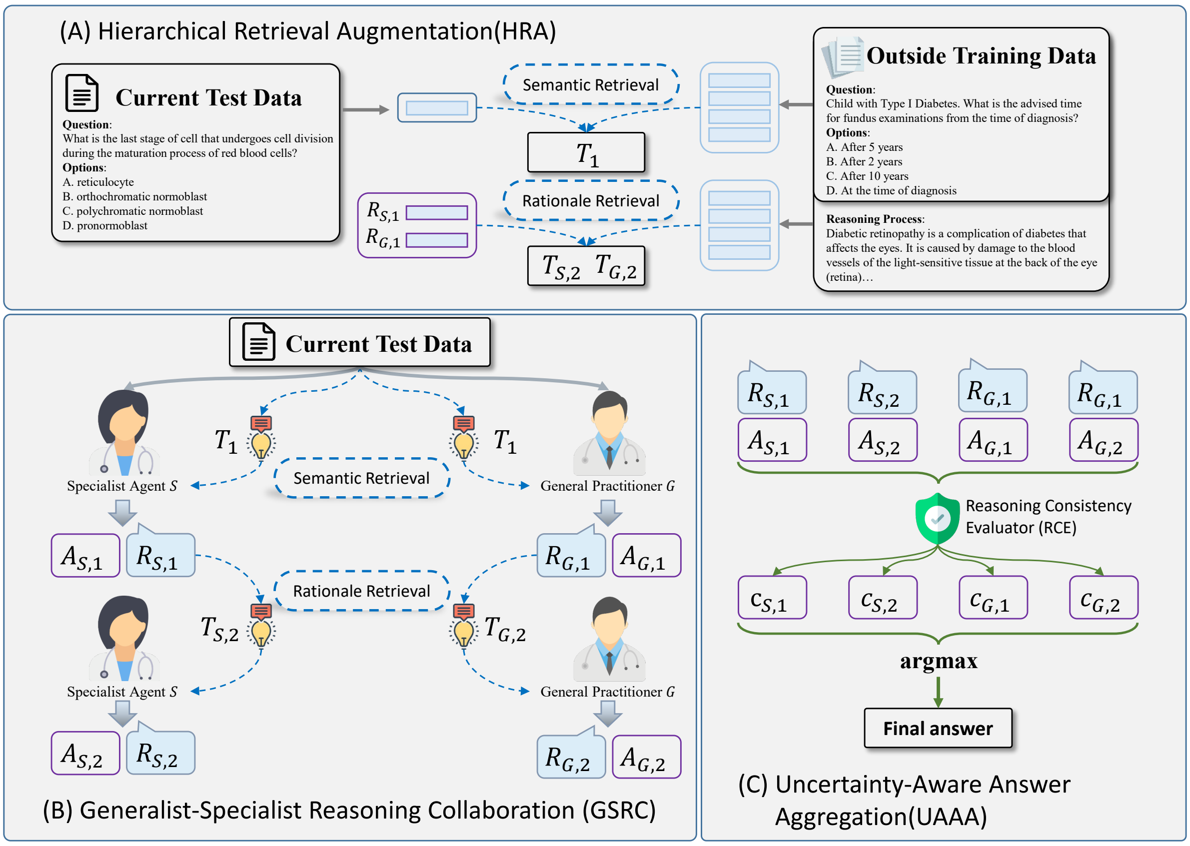

Figure 1: Overview of the proposed TAGS framework. The architecture consists of three modules: (A) **HRA (Hierarchical Retrieval Augmentation)**, a two-stage retrieval process that selects semantically relevant exemplars ($\mathcal{T}_1$) and refines them based on rationale alignment ($\mathcal{T}_{G,2}$, $\mathcal{T}_{S,2}$). (B) **GSRC (Generalist-Specialist Reasoning Collaboration)** employs dual-agent reasoning across two rounds, generating four candidate (Rationale, Answer) pairs. (C) **UAAA (Uncertainty-Aware Answer Aggregation)** assesses rationale consistency using the RCE and aggregates reliability scores (c) to determine the final answer.

at the reasoning level. Our framework addresses this limitation by employing a hierarchical retrieval strategy: first retrieving question-option exemplars, then refining based on CoT similarity. This enables alignment in both problem context and reasoning structure, thereby improving downstream multi-agent reasoning.

2.3 Multi-Agent Systems for Reasoning

Multi-agent systems (MAS) have emerged as a promising approach to enhance the robustness, diversity, and reliability of reasoning in complex tasks, including medical question answering. By orchestrating multiple reasoning paths or personas, MAS frameworks aim to mitigate biases and capture complementary perspectives, which are particularly critical in high-stakes medical decision-making. Recent studies have explored various MAS paradigms for medical reasoning. MedAgents (Tang et al., 2023) proposes a collaborative multi-agent framework where multiple agents independently generate answers and a majority voting scheme determines the final prediction. MDAgents (Kim et al., 2024a) further enhances this idea by introducing dynamic collaboration and adaptive feedback mechanisms among agents during the reasoning process. MedPrompt (Chen et al., 2024d) adopts a multi-round prompting strategy combined with ensemble voting to improve medical QA performance. Additionally, frameworks like Multi-Persona (Wang et al., 2023) and Self-Refine (Madaan et al., 2024) leverage self-collaboration and iterative self-feedback to strengthen individual agent reasoning capabilities. While multi-agent collaboration has demonstrated effectiveness in improving answer quality, it also introduces notable challenges. As highlighted in recent evaluations (Tang et al., 2023), excessive agent interactions may lead to reasoning conflicts, unstable decision paths, and increased inference costs. Recent studies (Cemri et al., 2025) further reveal that over-complex MAS architectures often suffer from systemic failures, such as miscommunication, vague role specification, and weak verification. Moreover, most existing MAS frameworks lack explicit mechanisms to assess the internal consistency between generated reasoning and final answers, which can limit reliability in clinical contexts. To address these limitations, we propose a lightweight General-Specialist Reasoning Collaboration (GSRC) strategy that pairs a generalist and a specialist agent in a complementary manner, promoting stable and robust medical reasoning with minimal inter-agent conflicts.

3 Methodology

We propose TAGS (Test-time Generalist–Specialist Reasoning with Retrieval-Augmentation and Uncertainty-Aware Verification), a parameter-efficient framework for medical question answering that operates entirely during inference. At its core is the Generalist–Specialist Reasoning Collaboration (GSRC), a dual-agent design that promotes reasoning diversity and domain alignment without requiring any parameter updates. To support GSRC, we introduce two auxiliary modules: Hierarchical Retrieval Augmentation (HRA), which supplies diverse and rationale-aligned exemplars, and Uncertainty-Aware Answer Aggregation (UAAA), which selects the final answer by evaluating the consistency of each reasoning path. As shown in Figure 1, these components form an integrated pipeline that enables robust, zero-shot clinical QA without model fine-tuning.

3.1 Hierarchical Retrieval Augmentation

Hierarchical Retrieval Augmentation (HRA) grounds reasoning in up-to-date evidence while injecting diverse paths for chain-of-thought (CoT) generation through a two-stage retrieval scheme. We retrieve from a frozen medical-QA corpus $\mathcal{D}$ whose entries are $d_i = (Q_i, O_i, A_i, R_i)$, where R_i denotes the CoT rationale. We use a frozen text encoder $\mathcal{E}(\cdot)$ based on M3-Embedding (Chen et al., 2024b), with 1024-dimensional output.

Stage 1: Initial semantic retrieval. We begin by embedding the query using a frozen encoder. Let $\mathbf{z} = \mathcal{E}(Q \oplus O)$, where the question Q and its options O are concatenated in standard order (A, B, C, D). Cosine similarity is computed against all candidate embeddings $\mathcal{E}(Q_i \oplus O_i)$ in the corpus. The top-K retrieved examples form:

$$\mathcal{T}_1 = \text{Top-K}\big\{d_i \in \mathcal{D} : \text{sim}(\mathbf{z}, \mathcal{E}(Q_i \oplus O_i))\big\}. \quad (1)$$

Stage 2: Rationale-guided retrieval. After Round-1 reasoning yields preliminary rationales $R_{G,1}$ and $R_{S,1}$, we retrieve exemplars whose stored rationales best match these CoTs:

$$\mathcal{T}_{G,2} = \text{Top-K}\big\{d_i \in \mathcal{D} : \text{sim}(\mathbf{r}_G, \mathcal{E}(R_i))\big\}, \quad (2)$$

$$\mathcal{T}_{S,2} = \text{Top-K}\big\{d_i \in \mathcal{D} : \text{sim}(\mathbf{r}_S, \mathcal{E}(R_i))\big\}. \quad (3)$$

By aligning on reasoning paths rather than surface form, Stage 2 injects complementary evidence beyond surface similarity, reducing the limitations of purely semantic matching.

3.2 Generalist–Specialist Collaboration

Given the retrieved exemplar sets from HRA, Generalist–Specialist Reasoning Collaboration (GSRC) performs dual-agent inference in two rounds by coupling broad medical knowledge with focused domain expertise. The system consists of a generalist agent $\mathcal{G}$ and a specialist agent $\mathcal{S}$, both instantiated as prompted roles of the same frozen LLM without parameter updates.

An auxiliary LLM role first infers the medical specialty most relevant to the query (Q, O), yielding a label s (e.g., `cardiology`). This label is then injected into the prompt for $\mathcal{S}$ as "You are a medical specialist in the field of [s]", guiding its reasoning toward domain-specific knowledge while preserving the core semantics of (Q, O). Further details are provided in Appendix B. The collaboration unfolds in two rounds that iteratively refine rationales and answers.

Round 1: Initial hypothesis generation. Both agents receive the query (Q, O) together with the semantically retrieved set $\mathcal{T}_1$ (§3.1); the specialist additionally sees the inferred specialty s. Each agent produces an initial CoT and answer:

$$\begin{aligned}(R_{G,1}, A_{G,1}) &= \mathcal{G}(Q, O, \mathcal{T}_1),\\ (R_{S,1}, A_{S,1}) &= \mathcal{S}(Q, O, \mathcal{T}_1, s).\end{aligned} \quad (4)$$

These preliminary CoTs trigger Stage 2 of HRA, which returns the rationale-aligned exemplar sets $\mathcal{T}_{G,2}$ and $\mathcal{T}_{S,2}$.

Round 2: Refined reasoning with aligned exemplars. Using the tailored sets, the agents generate updated rationales and answers:

$$\begin{aligned}(R_{G,2}, A_{G,2}) &= \mathcal{G}(Q, O, \mathcal{T}_{G,2}),\\ (R_{S,2}, A_{S,2}) &= \mathcal{S}(Q, O, \mathcal{T}_{S,2}, s).\end{aligned} \quad (5)$$

Finally, the four *(rationale, answer)* pairs are gathered into a candidate set

$$\mathcal{C} = \{ (R_{k,r}, A_{k,r}) \mid k \in \{G, S\},\ r \in \{1, 2\} \}, \quad (6)$$

which is then forwarded to the Uncertainty-Aware Answer Aggregation module (§3.3) for scoring and final selection.

3.3 Uncertainty-Aware Answer Aggregation

Uncertainty-Aware Answer Aggregation (UAAA) takes as input the candidate set $\mathcal{C}$ generated by GSRC (§3.2) and selects a single high-confidence answer through consistency-based scoring. To accomplish this, we define a *Reasoning Consistency Evaluator* (RCE), implemented as a separate zero-shot role of the same frozen LLM.

Given a candidate pair (R_k, A_k), the RCE assesses how well the rationale supports the answer in the context of the original query (Q, O), and assigns an integer score $c_k \in [0, 5]$, where higher values indicate stronger logical and clinical coherence. The scoring rubric is detailed in Appendix C. The final answer is selected as:

$$A_{\text{final}} = A_{k^\star}, \quad k^\star = \arg\max_{k \in \mathcal{C}} c_k. \quad (7)$$

In the case of ties, preference is resolved deterministically in the following order: specialist round 2, generalist round 2, specialist round 1, and generalist round 1. By explicitly verifying the internal consistency of each reasoning path, UAAA mitigates hallucination propagation and stabilizes final predictions, all without any parameter updates.

4 Experiments

4.1 Experimental Setting

Retrieval Dataset. We use the MedReason dataset (Wu et al., 2025) as our external retrieval corpus. It contains 32,682 medical QA pairs with clinically validated, step-by-step explanations generated via a knowledge graph–guided pipeline based on PrimeKG (Chandak et al., 2023). Unlike general CoT datasets, MedReason ensures factual correctness by filtering out chains that do not lead to the correct answer. We treat it as a structured knowledge base for retrieving semantically or logically relevant examples at inference. Dataset construction details and examples are provided in Appendix D.

Test Datasets. We evaluate TAGS on a curated benchmark of nine medical QA datasets selected from the MEDAGENTSBENCH framework (Tang et al., 2025), designed to assess complex clinical reasoning. The benchmark includes challenging subsets from: **MedQA** (Jin et al., 2021), a multilingual board-exam dataset (e.g., USMLE); **PubMedQA** (Jin et al., 2019), derived from biomedical literature with yes/no/maybe answers; **MedMCQA** (Pal et al., 2022), covering 21 medical subjects from Indian medical exams; **MedBullets** (Chen et al., 2024a), featuring long-context clinical questions; **MedExQA** (Kim et al., 2024b), emphasizing explainable QA across five specialties; **MedXpertQA** (Zuo et al., 2025), with subsets targeting reasoning and understanding; **MMLU** (Hendrycks et al., 2020) and **MMLU-Pro** (Wang et al., 2024), general benchmarks with medical subfields.

To better reflect real-world difficulty, we follow the hard subset construction pipeline proposed by MEDAGENTSBENCH. Questions are selected based on model failure rates (<50% accuracy across a set of strong models), medical topic coverage, and reasoning depth. Specifically, we include 100 hard questions each from MedQA, PubMedQA, MedMCQA, MedExQA, and MMLU-Pro; 100 from each MedXpertQA subset (Reasoning and Understanding); 89 from MedBullets; and 73 from MMLU. This results in a total of 862 expert-verified instances designed to stress-test the reasoning capabilities of large language models.

Baselines. We first compare our method against several widely adopted prompting and reasoning strategies that do not involve model updates: **(1) CoT (Chain-of-Thought)**(Wei et al., 2022): A prompting technique that guides the model to articulate intermediate reasoning steps before producing a final answer. **(2) CoT-SC (Chain-of-Thought with Self-Consistency)**(Wang et al., 2022): An extension of CoT that generates multiple reasoning paths and selects the most consistent answer via majority voting. **(3) Multi-Persona**(Wang et al., 2023): A method that simulates multiple expert personas to collaboratively reason through clinical questions. **(4) Self-Refine**(Madaan et al., 2024): A self-improvement framework in which the model iteratively refines its own responses across multiple reasoning stages. **(5) MedAgents**(Tang et al., 2023): A domain-specific multi-agent framework that employs multiple specialist agents for collaborative clinical reasoning. **(6) MDAgents**(Kim et al., 2024a): A lightweight variant of MedAgents that combines minimal agent collaboration

Method	MedQA	PubMedQA	MedMCQA	MedBullets	MMLU	MMLU-Pro	MedExQA	MedXpert-R	MedXpert-U	Average
GPT-4o	32.0	9.0	25.0	19.1	24.7	21.0	18.0	7.0	6.0	18.0
+ few-shot	31.0	16.0	34.0	16.9	32.9	27.0	17.0	8.0	11.0	21.7
+ RAG	42.0	12.0	30.0	22.5	20.5	37.0	15.0	19.0	10.0	23.1
+ CoT	39.0	10.0	30.0	28.1	26.0	35.0	24.0	12.0	15.0	24.3
+ CoT-SC	37.0	6.0	35.0	30.3	30.1	43.0	22.0	10.0	14.0	25.3
+ Multi-Persona	45.0	15.0	25.0	29.2	37.0	42.0	21.0	10.0	16.0	26.7
+ Self-Refine	41.0	13.0	34.0	28.1	34.2	34.0	22.0	17.0	19.0	26.9
+ MedAgents	43.0	15.0	30.0	27.0	28.8	8.0	19.0	3.0	6.0	20.0
+ MDAgents	36.0	11.0	22.0	21.3	24.7	8.0	13.0	4.0	5.0	16.1
+ MedPrompt	34.0	11.0	26.0	22.5	26.0	22.0	16.0	14.0	9.0	20.1
+ Ours	54.0	13.0	32.0	33.7	45.2	47.0	17.0	22.0	22.0	31.8

Table 1: Performance heatmap by methods and datasets. All tasks are evaluated on the HARD set with Pass@1 Accuracy (%) using GPT-4o base model.

Method	MedQA	PubMedQA	MedMCQA	MedBullets	MMLU	MMLU-Pro	MedExQA	MedXpert-R	MedXpert-U	Average
DeepSeek-R1	38.0	11.0	28.0	36.0	32.9	36.0	20.0	20.0	23.0	27.2
+ few-shot	27.0	12.0	32.0	33.7	35.6	41.0	27.0	11.0	9.0	25.4
+ RAG	49.0	20.0	31.0	43.8	53.8	42.0	25.0	28.0	26.0	35.4
+ CoT	47.0	12.0	31.0	39.3	38.4	35.0	22.0	27.0	27.0	31.0
+ CoT-SC	52.0	14.0	32.0	43.8	45.2	38.0	24.0	17.0	26.0	32.4
+ Multi-Persona	52.0	18.0	37.0	42.7	42.5	38.0	26.0	23.0	26.0	33.9
+ Self-Refine	33.0	17.0	30.0	34.8	27.4	22.0	24.0	12.0	13.0	23.7
+ MedAgents	48.0	21.0	22.0	44.9	43.8	35.0	27.0	22.0	25.0	32.1
+ MedPrompt	46.0	14.0	30.0	38.2	45.2	27.0	24.0	8.0	7.0	26.6
+ Ours	55.0	28.0	35.0	52.8	61.6	53.0	26.0	36.0	49.0	44.0

Table 2: Performance heatmap by methods and datasets. All tasks are evaluated on the HARD set with Pass@1 Accuracy (%) using DeepSeek-R1 base model.

with retrieval augmentation to improve reasoning. **(7) MedPrompt** (Chen et al., 2024d): A retrieval-augmented prompting strategy that integrates semantically similar historical cases to enhance clinical inference.

We additionally report results under a **few-shot** baseline, where five training examples from the target dataset are retrieved and used as in-context demonstrations for single-pass inference. We also include a **RAG** baseline, which retrieves the top-K most semantically similar questions with accompanying CoTs from the MedReason dataset and feeds them directly to the model. This RAG setting shares the same retrieval setup but excludes agent collaboration and verification, highlighting the value of structured reasoning.

We further evaluate our method against several strong open-source foundation models and their medically adapted variants: **(1) Qwen2.5-7B** (Yang et al., 2024a): A 7B general-purpose model instruction-tuned for diverse tasks, evaluated both with and without CoT prompting. **(2) LLaMA-3-8B** (Grattafiori et al., 2024): Meta's latest 8B instruction-tuned model with improved reasoning capabilities. **(3) HuatuoGPT-o1-7B** (Chen et al., 2024c): A 7B model fine-tuned for complex medical reasoning via reinforcement learning. **(4) HuatuoGPT-o1-8B** (Chen et al., 2024c): An enhanced 8B version of HuatuoGPT, optimized for clinical inference tasks. **(5) MedLLaMA-3-8B-v1.0** (Qiu et al., 2024): A medical-adapted variant of LLaMA-3 trained on biomedical corpora. **(6) MedLLaMA-3-8B-v2.0**: An updated release with improved performance on expert-level medical benchmarks. **(7) OpenBioLLM-8B** (Pal and Sankarasubbu, 2024): An open-source 8B biomedical language model fine-tuned for healthcare and life sciences applications.

Evaluation Metrics Following (Tang et al., 2023), we report **Pass@1 Accuracy** as the evaluation metric, which measures whether the model's first generated answer exactly matches the ground-truth answer.

Reproducibility. All experiments were conducted using Python 3.10 and PyTorch 2.4.0 on four NVIDIA H100 GPUs, each with 80 GB of memory. For proprietary LLM baselines such as GPT-4o and DeepSeek-R1, we accessed the models through their official APIs and ensured consistent use of the same model version across all runs. For open-source models, we directly loaded checkpoint weights from their respective official Hugging Face repositories to ensure reproducibility and transparency.

Method	MedQA	PubMedQA	MedMCQA	MedBullets	MMLU	MMLU-Pro	MedExQA	MedXpert-R	MedXpert-U	Average
Qwen2.5-7B	16.0	16.0	24.0	4.5	13.7	26.0	9.0	10.0	8.0	14.1
Llama-3-8B	18.0	13.0	23.0	16.9	11.0	23.0	11.0	10.0	4.0	14.4
HuatuoGPT-o1-7B	22.0	21.0	26.0	12.4	13.7	29.0	9.0	11.0	7.0	16.8
HuatuoGPT-o1-8B	29.0	20.0	33.0	20.2	21.9	17.0	18.0	16.0	7.0	20.2
MedLlama-3-8B-v1.0	24.0	20.0	22.0	14.6	16.4	12.0	12.0	11.0	11.0	15.9
MedLlama-3-8B-v2.0	28.0	25.0	22.0	22.5	32.9	12.0	22.0	9.0	8.0	20.2
OpenBioLLM-8B	19.0	29.0	20.0	19.1	21.9	2.0	17.0	7.0	3.0	15.3
Qwen2.5-7B + Ours	28.0	25.0	24.0	14.6	35.6	25.0	16.0	18.0	29.0	23.9

Table 3: Comparison with fine-tuned medical LLMs on nine MedQA benchmarks.

4.2 Compared with Prompting and MAS

We evaluate the effectiveness of TAGS by comparing it with a diverse set of prompting-based and multi-agent reasoning baselines across nine challenging MedQA benchmarks. Tables 1 and 2 summarize the results in terms of **Pass@1 Accuracy**, evaluated on the HARD split using two foundational LLMs: GPT-4o and DeepSeek-R1. Under the GPT-4o setting, TAGS achieves the highest average accuracy of 31.8%, outperforming all baselines including Self-Refine (26.9%), MedAgents (20.0%), and CoT-SC (25.3%). The most notable improvements appear on MedQA (+9.0 over Multi-Persona), MMLU (+8.2 over Multi-Persona), and MedXpert-R (+3.0 over RAG), highlighting the impact of verifier-guided aggregation and structured multi-agent reasoning. TAGS also surpasses standard few-shot and RAG baselines by margins of +10.1 and +8.7 respectively. With the DeepSeek-R1 base model, TAGS achieves an average accuracy of 44.0%, outperforming CoT-SC (32.4%), Multi-Persona (33.9%), and MedAgents (32.1%). Notably, MDAgents consistently failed under this setting due to format inconsistencies. TAGS also surpasses the few-shot and RAG baselines by margins of +18.6 and +8.5, respectively, demonstrating the scalability of our framework across both general and domain-specific tasks.

4.3 Compared with Fine-Tuned LLMs

To further contextualize the performance of our TAGS framework, we evaluate its effectiveness when integrated with the Qwen2.5-7B base model and compare its performance against a series of prominent open-source and medically fine-tuned large language models across the same nine challenging MedQA datasets. The results of this comparison are presented in Table 3. As shown in the table, our TAGS framework substantially boosts the zero-shot question answering capability of the base Qwen2.5-7B model, improving its average accuracy from 14.1% to 23.9%. In particular, TAGS demonstrates robust performance gains on difficult benchmarks such as MedQA (+12.0 percentage points), MMLU (+21.9 percentage points), and MedXpert-U (+21.0 percentage points). Notably, our inference-only strategy even outperforms several models that have been fine-tuned with domain-specific medical corpora or expert feedback, such as MedLLaMA-3-8B and the HuatuoGPT-o1 variants, on the majority of the evaluated datasets. These results strongly highlight the significant potential of structured retrieval and multi-agent reasoning, combined with uncertainty-aware verification, to effectively close the performance gap with models requiring extensive fine-tuning, while retaining the inherent flexibility and adaptability of a zero-shot approach.

4.4 Ablation Study

We conduct an ablation study on the Qwen2.5-7B model to assess the impact of each module in the TAGS framework, focusing on two datasets: MMLU and MedXpert-U. As shown in Table 4, the baseline Qwen2.5-7B model without any module achieves 13.7% and 8.0% accuracy on MMLU and MedXpert-U, respectively. Incorporating retrieval augmentation (RAG) improves the scores to 20.5% and 10.0%, highlighting the utility of external knowledge retrieval. Adding the generalist (G) and specialist (S) agents with a majority voting mechanism raises MMLU accuracy to 30.1%, while MedXpert-U remains at 10.0%, indicating limited impact on domain-specific tasks. Integrating hierarchical retrieval augmentation (HRA) results in notable gains, with 34.2% for the generalist and 32.9% for the specialist, indicating the value of reasoning-guided retrieval. The complete TAGS framework, including uncertainty-aware aggregation (UAAA), achieves the best performance of

RAG	$\mathcal{G}$	$\mathcal{S}$	HRA	UAAA	MMLU	MedXpert-U
					13.7	8.0
✓					20.5	10.0
✓	✓	✓			30.1	10.0
✓	✓		✓	✓	34.2	16.0
✓		✓	✓	✓	32.9	18.0
✓	✓	✓	✓		31.5	22.0
✓	✓	✓	✓	✓	35.6	29.0
w/o-top10	✓	✓	✓	✓	37.0	24.0
✓	✓	w-3rd	✓	✓	34.2	23.0

Table 4: Ablation Study on TAGS using Qwen2.5-7B. RAG: Retrieval-Augmented Generation; $\mathcal{G}$: Generalist Agent; $\mathcal{S}$: Specialist Agent; HRA: Hierarchical Retrieval Augmentation; UAAA: Uncertainty-Aware Answer Aggregation.

35.6% and 29.0%, confirming the synergistic impact of structured retrieval, dual-agent reasoning, and verification.

To assess whether our framework relies on retrieving semantically closest examples, we introduce **RAG-w/o-topk**, which explicitly excludes the top-10 most similar questions during retrieval. Despite this restriction, performance only marginally decreases, demonstrating that our model benefits primarily from exposure to valid reasoning patterns rather than from copying specific answers. We further introduce **$\mathcal{S}$-w-3rd**, a variant that assigns the 3rd most relevant specialist instead of the top-ranked one. This setting simulates scenarios where the domain classifier misidentifies the optimal expert, which may occur in real-world deployments. The performance drops only slightly under this perturbation, suggesting that TAGS does not strongly depend on perfect specialist selection. Even sub-optimal specialists can provide useful guidance, highlighting the robustness of our framework.

4.5 Hyperparameter Analysis

Figure 2 shows TAGS' sensitivity to two key hyper-parameters: the number of specialist agents and the retrieval size K. In Figure 2(a), adding one specialist to the generalist improves accuracy from 34.2% to 45.2% on MMLU and from 16.0% to 22.0% on MedXpert-U. However, adding more specialists brings limited or no further gains, likely due to redundancy or conflicts in reasoning paths. Figure 2(b) shows that accuracy peaks at $K = 2$ and declines with larger K, as additional exemplars may introduce noise or irrelevant content that misleads the model. These results support our choice of using one specialist and $K = 2$ as the default configuration, balancing diversity and robustness.

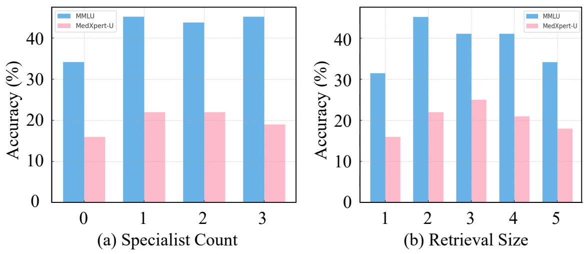

Figure 2: Hyper-parameter sensitivity analysis of specialist count and retrieval size in relation to accuracy.

4.6 Inference Efficiency

On the MedQA dataset with GPT-4o, TAGS takes 72 seconds per question on average, which is longer than CoT-SC (27.7s) but shorter than Multi-Persona (109.6s). Although slower than simple prompting, TAGS achieves substantially higher accuracy. Both reasoning and verification are parallelizable, enabling efficient deployment in real-world clinical settings. This moderate inference cost represents a favorable trade-off for improved robustness and reliability. Additionally, these stages can be parallelized across GPU streams or executed via asynchronous API calls to further speed up inference.

5 Conclusion

We presented TAGS, a parameter-efficient and test-time-only framework designed to enhance reliability in medical question answering without model fine-tuning. TAGS introduces a structured reasoning paradigm through generalist–specialist reasoning collaboration, which combines the breadth of a generalist with the depth of a specialist to generate complementary inference paths. This collaboration is guided by hierarchical retrieval augmentation, which retrieves exemplars at both semantic and rationale levels to enrich reasoning diversity, and finalized by uncertainty-aware answer aggregation to select robust answers. Extensive experiments on nine challenging MedQA benchmarks, spanning general-purpose and fine-tuned LLMs, demonstrate TAGS' consistent superiority over prompting-based, retrieval-augmented, and multi-agent baselines. Notably, our method delivers substantial improvements even for compact 7B models, highlighting its adaptability across model scales. TAGS offers a practical, inference-only alternative for trustworthy medical AI and opens promising directions for adaptive retrieval, dynamic agent collaboration, and scaling to multimodal or real-world clinical QA workflows.

Limitations

While TAGS offers a robust, inference-only approach to medical QA, it carries several limitations. First, it depends heavily on the coverage and quality of the external retrieval corpus: gaps or biases in the QA database may lead to missing or misleading exemplars, particularly for rare diseases or newly emerging clinical scenarios. Second, the Reasoning Consistency Evaluator (RCE) is itself a zero-shot LLM prompt and may inherit the same hallucination tendencies or biases as the generator agents, potentially mis-scoring perfectly valid but unconventional reasoning chains. Third, the two-round retrieval and dual-agent design, while effective, substantially increases inference latency and API cost compared to single-pass prompting; this may limit real-time deployment in resource-constrained clinical settings.

Additionally, our current evaluation focuses solely on answer accuracy (Pass@1), without assessing the interpretability or faithfulness of reasoning paths. Future work may benefit from human evaluation or rationale consistency metrics to further assess clinical applicability.

Moreover, our specialty inference component can occasionally misclassify the most relevant domain, which, although gracefully handled, may still introduce suboptimal reasoning contexts. Finally, our evaluation is confined to English-language, multiple-choice benchmarks and does not cover open-ended clinical dialogs, multimodal data (e.g., images, lab reports), or non-English patient populations. Addressing these limitations will require enriching and updating the retrieval corpus, developing more calibrated or human-in-the-loop verifier mechanisms, optimizing retrieval budgets and round counts, and extending evaluation to diverse, real-world clinical workflows.

References

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. 2025. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335.

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

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

Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. 2023a. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv e-prints*, pages arXiv–2305.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. 2023b. A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. 2025b. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Xuhang Chen, Shenghong Luo, Chi-Man Pun, and Shuqiang Wang. 2024d. MedPrompt: Cross-modal prompting for multi-task medical image translation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 61–75. Springer.

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.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Qiangqiang He, Shuwei Qian, Jie Zhang, and Chongjun Wang. 2025. Inference retrieval-augmented multimodal chain-of-thoughts reasoning for language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

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*.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024a. MDAgents: An adaptive collaboration of llms in medical decision making. *arXiv preprint arXiv:2404.15155*.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024b. MedExQA: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

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.

Malaikannan Sankarasubbu Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. *Hugging Face repository*.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May D Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. *arXiv preprint arXiv:2405.03000*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. MedAgents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al.

2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. 2022. Prompting decision transformer for few-shot policy generalization. In *international conference on machine learning*, pages 24631–24645. PMLR.

Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, et al. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*.

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

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1028–1043.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

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

A Ethics Statement

This work relies solely on publicly available medical question answering datasets, including MedQA, PubMedQA, MedMCQA, and others curated in the MEDAGENTSBENCH framework (Tang et al., 2025). These datasets are de-identified and collected from open educational or biomedical sources such as medical board exams and peer-reviewed literature. No private health records or patient-identifiable information were used.

Our proposed framework operates entirely at test time and does not require any model fine-tuning or user data collection. All evaluations are conducted offline on benchmark datasets, and no deployment in real clinical settings has been performed.

While our method is designed to improve the robustness and reliability of medical LLMs, it is not intended for use in high-stakes clinical decision-making without appropriate human oversight. We emphasize that the generated answers should not be interpreted as medical advice. Future work may involve incorporating human-in-the-loop mechanisms and broader impact assessments before real-world deployment.

B Generalist–Specialist System Roles and Prompt Templates

B.1 System Prompt for Specialist and Generalist

To ensure consistency and clarity across different model roles, we define structured system prompts tailored to each classifier in our multi-agent framework. These prompts specify role-specific reasoning strategies and output formats, enabling the models to adopt appropriate clinical reasoning behaviors under zero-shot test-time conditions.

The system prompt for the specialist categorization agent is presented in Table 5, while the diagnostic prompt for the specialist agent is shown in Table 6. The prompt for the generalist agent is provided in Table 7.

B.2 Prompt Organization and Structure

To ensure faithful and consistent model behavior across different roles and stages of inference, we design modular and task-specific prompt templates. These templates guide the models in both few-shot reasoning and auxiliary classification tasks.

Specifically, the specialist classification prompt (Table 8) is used to determine the relevant subfields

Specialist Categorization — System Prompt

You are a senior medical expert tasked with classifying clinical multiple-choice problems into the most relevant areas of medical science.

Your role is strictly to determine and output the classification.

Important: *Do not provide any explanation, reasoning, or commentary. Only output the final classification strictly following the format.*

Table 5: System prompt for the specialist categorization.

Specialist Agent — System Prompt

You are an experienced specialist in `{domain}`. Your role is to carefully analyze clinical multiple-choice questions from the standpoint of a `{domain.lower()}` expert. You should reason by focusing on the interpretation of symptoms, underlying pathophysiology, and domain-specific diagnostic principles.

First, review the provided reference examples and understand their reasoning patterns.
Then, based on your specialist knowledge, perform *structured, step-by-step* reasoning for the new question.

Required output format

```
Thought: [your detailed step-by-step reasoning]
Answer: [one of A, B, C, . . . ]
```

Table 6: System prompt for the specialist agent.

Generalist Agent — System Prompt

You are a general practitioner trained to manage a wide range of clinical conditions. Your task is to evaluate clinical multiple-choice questions using broad, cross-disciplinary medical knowledge. Focus on extracting key clinical findings, ruling out unlikely diagnoses, and applying general reasoning principles.

First, analyze the reference examples to understand their diagnostic thought process.
Then, produce a *step-by-step* analysis for the new question.

Required output format

```
Thought: [your detailed step-by-step reasoning]
Answer: [one of A, B, C, . . . ]
```

Table 7: System prompt for the generalist agent.

of medicine needed to solve a given question, serving as a basis for downstream role assignment and retrieval. Meanwhile, the few-shot prompt template (Table 9) provides structured instructions and reference examples to facilitate reasoning transfer for clinical question answering.

Specialist Classifier — Prompt

Task Instructions

- Carefully analyze the following medical question:

  `'''{question}'''`

- The corresponding options are:

  `'''{options}'''`

- Based on both the question and the options, determine the top **3 most relevant subfields of medicine** that are required to solve this question.
- You must only output in the exact format:

  `Medical Field: Field1 | Field2 | Field3`

Table 8: Prompt used for classifying medical questions into relevant specialist subfields.

C The Reasoning Consistency Evaluator Rubric and Prompt

To robustly aggregate multi-agent responses, we introduce a reliability scoring mechanism that evaluates the consistency between an agent's reasoning and its final answer. In scenarios where a question has many answer options (e.g., N), simple majority voting becomes inefficient — achieving a reliable consensus typically requires at least $N+1$ agreeing agents.

To address this, we employ a scoring-based verification strategy: each agent's reasoning is evaluated by a separate verifier agent that assigns a reliability score ranging from 1 to 5. This enables us to treat scores as soft confidence signals and aggregate responses more efficiently, even when only a few answers are available. The resulting per-sample reliability sum lies in the range of 4–20 (with 4 verifiers), providing fine-grained guidance for final answer selection. The full scoring prompt is shown in Table 10.

D Reference CoT Dataset Examples

We adopt the MedReason dataset (Wu et al., 2025) as our external reference corpus to support

Few-shot Prompt Template

Header
"Your task is to solve the following clinical multiple-choice question."

Question Block
Present the target question text, then list answer options (A/B/C/...).

Instruction Block

- The prompt shows N solved *reference examples*.
- Each example contains:
  - **Finding Reasoning Paths**: brainstorming approaches
  - **Reasoning Process**: a structured, step-by-step solution
- Focus *only* on learning the reasoning patterns.
- For the *new* question, generate your **own** reasoning and answer.
- Constraints:
  - Always choose one of the provided options—never "unknown"
  - Follow the exact output format shown below

Output Format Hint
```
Thought: [your detailed step-by-step reasoning]
Answer: [one of A, B, C, . . . ]
```

Reference Examples
For each retrieved example i insert:
```
Reference Example i:
<question text>
<options>
Thought: <reference rationale>
```

Table 9: Prompt template used for few-shot reasoning with retrieved reference examples.

Reliability Scoring — System Prompt

A clinical AI agent has answered the following multiple-choice question:

Question:
"`{question}`"

Options:
"`{$options_str$}`"

The agent provided the following reasoning:
Thought:
"`{thought}`"

Final Answer:
"`{answer}`"

Your Role:
You are a critical-thinking medical reviewer. Your task is to assign a reliability score from 1 to 5 based on how well the reasoning supports the answer.

Scoring Scale:

- 5 — Reasoning is complete, medically accurate, and fully supports the answer.
- 4 — Mostly correct with minor issues, but the answer is still justified.
- 3 — Reasoning has some issues or omissions, but partially supports the answer.
- 2 — Reasoning is flawed or incomplete; answer is weakly supported.
- 1 — Reasoning is incorrect or misleading; answer is not justified.

Output Format:
`Score: [1-5]`

Table 10: System prompt used by the reasoning consistency evaluator to assign a soft reliability score.

retrieval-augmented reasoning. MedReason comprises 32,682 high-quality question–answer pairs, each accompanied by detailed, clinically grounded chain-of-thought (CoT) explanations. The dataset is constructed through a knowledge graph–guided pipeline that ensures both logical consistency and medical factuality.

Specifically, the authors first collect QA pairs from seven public medical benchmarks, including MedQA, MedMCQA, PubMedQA, MMLU, MedXpert, Huatuo, and HLE. For each QA pair, relevant medical entities are extracted from both the question and the answer using GPT-4o and are then mapped to a structured medical knowledge graph, PrimeKG. Next, the shortest reasoning paths connecting the question and answer entities within the graph are retrieved and pruned using LLM-based selection to retain only clinically relevant paths. These paths serve as scaffolds for guiding step-by-step CoT generation.

To guarantee data quality, each generated reasoning trace is verified by prompting the LLM to reproduce the original answer solely based on the CoT explanation. If the answer cannot be recovered, the CoT is discarded. This quality filtering process reduces 45K generated samples to a final dataset of 32,682 verified CoTs, ensuring that each retained example is both logically sound and clinically accurate.

The 7,168th indexed sample from MedReason is shown below as an example:

Question:
A young male patient presents with complete rectal prolapse. The surgery of choice is?

Answer Choices:
A. Abdominal rectopexy
B. Delerom's procedure

C. Anterior resection
D. Goodsall's procedure
Answer:
Abdominal rectopexy. Explanation: Surgery is required, and the operation can be performed the perineal or the abdominal approaches. An abdominal rectopexy has a lower rate of recurrence, as an abdominal procedure risks damage to the pelvic autonomic nerves, resulting in possible sexual dysfunction, a perineal approach is also usually preferred in young men. Ref: Bailey & Love's Short Practice of Surgery, 25 ed., p. 1225
Reasoning:
Finding reasoning paths: 1. Rectal prolapse (disease) -> Surgical treatment options -> Perineal approach vs. Abdominal approach 2. Abdominal approach -> Abdominal rectopexy -> Risk of recurrence vs. risk of complications 3. Abdominal rectopexy -> Potential complications -> Pelvic autonomic nerve damage -> Male sexual dysfunction

Reasoning Process: 1.**Understanding Rectal Prolapse and Treatment Options**: Rectal prolapse is a condition where the rectum protrudes through the anus. Surgical intervention is often required, and the two main approaches are the perineal and abdominal approaches.

2.**Evaluating Surgical Approaches**: -**Perineal Approach**: This is generally less invasive and may be preferred in older or frail patients due to lower surgical risk. However, it is associated with a higher rate of recurrence. -**Abdominal Approach**: This includes procedures like abdominal rectopexy, which tend to have a lower recurrence rate. However, it is more invasive and carries a risk of complications.

3.**Abdominal Rectopexy and Complications**: -**Recurrence Rate**: Abdominal rectopexy is favored for its lower recurrence rate compared to perineal procedures. -**Complications**: A significant risk associated with the abdominal approach is potential damage to the pelvic autonomic nerves. This can lead to complications such as male sexual dysfunction, which is an important consideration in a young male patient.

Conclusion: Abdominal rectopexy is the preferred surgical option for a young male patient with complete rectal prolapse due to its lower rate of recurrence. However, it is important to consider the risk of pelvic autonomic nerve damage, which could lead to male sexual dysfunction. This risk should be weighed against the benefits of a lower recurrence rate when deciding on the surgical approach.

E Test Dataset Examples

Test Datasets. We conduct experiments on a curated benchmark comprising challenging subsets of nine medical QA datasets, selected from the MEDAGENTSBENCH framework (Tang et al., 2025). This benchmark is specifically designed to evaluate complex medical reasoning and includes questions that require multi-step clinical inference, diagnosis formulation, and treatment planning. The original datasets include: **MedQA** (Jin et al., 2021), a multilingual dataset with English questions derived from professional medical board exams (e.g., USMLE), focusing on diagnosis and treatment knowledge. Each question includes four answer choices, with an average length of 167 tokens. **PubMedQA** (Jin et al., 2019), constructed from biomedical research articles in PubMed, requires reasoning over scientific evidence to determine yes/no/maybe answers. It contains long passages with an average length of 316 tokens. **MedMCQA** (Pal et al., 2022), sourced from Indian medical entrance exams (AIIMS and NEET-PG), comprises short multiple-choice questions across 21 medical subjects, averaging 18.7 tokens per question. **MedBullets** (Chen et al., 2024a), a high-yield medical education platform, provides detailed clinical questions with up to five answer options and longer contexts (avg. 213 tokens). **MedExQA** (Kim et al., 2024b), covers five medical specialties and emphasizes both correctness and explainability. It includes 935 questions, each with four options. **MedXpertQA** (Zuo et al., 2025), consists of two subsets: *Reasoning* and *Understanding*, specifically annotated to distinguish complex clinical inference tasks from knowledge recall. **MMLU** (Hendrycks et al., 2020) and **MMLU-Pro** (Wang et al., 2024), general-purpose multitask benchmarks that include medical subsets testing both factual recall and abstract reasoning.

To better reflect real-world difficulty, we follow the hard subset construction pipeline proposed by MEDAGENTSBENCH. Questions are selected based on model failure rates (<50% accuracy across a set of strong models), medical topic coverage, and reasoning depth. Specifically, we include 100 hard questions each from MedQA, PubMedQA, MedMCQA, MedExQA, and MMLU-Pro; 100 from each MedXpertQA subset (Reasoning and Understanding); 89 from MedBullets; and 73 from MMLU. This results in a total of 862 expert-verified instances designed to stress-test the reasoning capabilities of large language models. Representative examples from the nine datasets are provided in Table 11.

F Fail case of existing method

To illustrate the limitations of existing methods, we present a representative failure case from DeepSeek-R1 on a clinically complex MedQA question. **Scenario:** A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the procedure, the resident inadvertently cuts a flexor tendon, which is then repaired without complication. The attending advises the resident to exclude this incident from the operative report to avoid patient concern.

Question: *What is the correct next action for the resident to take?*

Answer Options:

- (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

Predicted Answer: (A) - Disclose the error and document it.

Model's Reasoning: <think> Okay, let's see. The question is about a resident who accidentally cuts a flexor tendon during carpal tunnel repair and the attending tells them not to report it. The options are A to D, and the generated solutions all say A.

First, the key points here are medical ethics, specifically honesty and patient disclosure. The resident made an error, even if it was repaired. The attending is advising to omit this from the operative report and not tell the patient to avoid worry.

In medical practice, complications should be documented and disclosed to the patient. The principle of informed consent requires patients to know about any complications, even if resolved. Failing to document could lead to legal issues and violates professional standards.

Option A says to disclose and document, which aligns with ethical guidelines. The other options either don't directly address informing the patient (like B, which is telling the attending but not necessarily documenting) or escalate without first handling it properly (C, reporting; D, refusing to dictate).

Since all solutions suggest A, the answer is likely A. They must inform the patient and include it in the report despite the attending's advice. </think>
A

Issue:
The model selected **(A)** as the answer, which is inconsistent with the reference solution **(B)**. The error occurred because the model focused solely on disclosure and documentation, ignoring the specific instruction to address the attending's unethical guidance.

Dataset	Question & Options & Answer
MedQA	**Question** A 39-year-old woman presents with headaches, constipation, increased thirst, and increased urinary frequency. Her blood pressure is 160/100 mm Hg and pulse is 67/min. Blood tests reveal high blood glucose and low potassium level. Abdominal CT scan shows a well-circumscribed homogeneous mass in the left adrenal gland. The hormone excess responsible for her symptoms uses which of the following pathways to exert its action? **Options** (A): "Intracellular receptors", (B): "cAMP pathway", (C): "cGMP pathway", (D): "JAK/STAT pathway". **Answer** (A)
PubMedQA	**Question** Sternal fractures in childhood are rare. The aim of the study was to investigate the accident mechanism, the detection of radiological and sonographical criteria and consideration of associated injuries.the period from January 2010 to December 2012 all inpatients and outpatients with sternal fractures were recorded according to the documentation.total of 4 children aged 5-14ŭ00a0years with a sternal fracture were treated in 2ŭ00a0years, 2ŭ00a0children were hospitalized for pain management and 2 remained in outpatient care.fracture in growing children : A rare and often overlooked fracture?**Options** (A): "yes", (B): "no", (C): "maybe". **Answer** (C)
MedMCQA	**Question** Minimum number of lobes require to form a tooth? **Options** (A): "1", (B): "2", (C): "3", (D): "4". **Answer** (C)
MedBullets	**Question** A 22-year-old woman presents to the emergency department with shortness of breath. She was hiking when she suddenly felt unable to breathe and had to take slow deep breaths to improve her symptoms. The patient is a Swedish foreign exchange student and does not speak any English. Her medical history and current medications are unknown. Her temperature is 99.5ŭ00b0F (37.5ŭ00b0C), blood pressure is 127/68 mmHg, pulse is 120/min, respirations are 22/min, and oxygen saturation is 90% on room air. Physical exam is notable for poor air movement bilaterally and tachycardia. The patient is started on treatment. Which of the following parameters including forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC), and diffusing capacity of carbon monoxide (DLCO) most appropriately describes this patient's underlying pathology? **Options** (A): "Decreased airway tone", (B): "Increased FEV1", (C): "Increased FEV1/FVC", (D): "Increased FVC", (E): "Normal DLCO". **Answer** (E)
MMLU	**Question** How many different types of microorganisms may colonize the mouth? **Options** (A): "35", (B): "100", (C): "350", (D): "500". **Answer** (C)
MMLU-Pro	**Question** How are new polyomaviruses detailed? **Options** (A): "Shot gun sequencing", (B): "Cultivation in human neural cells", (C): "Deep pyro sequencing (NGS)", (D): "Monoclonal antibody techniques". **Answer** (A)
MedExQA	**Question** Which biological tissue has the highest viscosity?**Options** (A): "Blood", (B): "Bone", (C): "Soft tissue", (D): "Water". **Answer** (B)
MedXpert-R	**Question** A 52-year-old paralegal comes to the clinic reporting chronic low back pain that has gradually worsened over two years. She describes an achy pain rated at 7/10 and spends considerable time at her computer desk. She experiences numbness, tingling, and radiating pain down her right leg during prolonged sitting or standing. Her examination reveals:- No urinary/bowel incontinence or perineal numbness- Right lower extremity strength of 4/5- Decreased sensation in right leg- Right patellar reflex grade 1/4- Positive straight leg raise test on right sideWhich nerve root levels are most likely involved in this presentation? **Options** (A) L1-L2 (B) L5-S1 (C) L2-L3 (D) S2-S3 (E) L4-L5 (F) L4-S1 (G) L3-L4 (H) S1-S2 (I) T12-L1 (J) L5-S2. **Answer** G
MedXpert-U	Which hypoxic cell radiosensitizer, known for undergoing redox recycling or decomposing into a toxic product, has demonstrated effectiveness in the treatment of head and neck cancer? **Options** (A): Doranidazole, (B): Tirapazamine, (C): Camptothecin, (D): Misonidazole, (E): Pimonidazole, (F): Nimorazole, (G): Sanazole, (H): Cetuximab, (I): Etanidazole, (J): Methotrexate. **Answer** (F)

Table 11: Dataset examples with corresponding questions and answer options from the nine test datasets.