

Trustworthy Medical LLMs via Bayesian Belief Updating

Mengyao Guo^{1*} Shufang Xie^{2*} Xiuying Chen³ Shuqi Li⁴ Hao Jiang⁵
Bo Du⁶ Rui Yan^{7†}

¹School of Information, Renmin University of China,

²Gaoling School of Artificial Intelligence, Renmin University of China, ³MBZUAI,

⁴KAUST, ⁵Beijing Chasing Stars Technology Co., Ltd.,

⁶School of Computer Science, Wuhan University,

⁷School of Artificial Intelligence, Wuhan University

{guomy7299, shufangxie}@ruc.edu.cn

Abstract

Large Language Models (LLMs) show strong potential in medical QA, but still suffer from hallucinations in multi-step clinical reasoning. Existing methods fail to assess confidence at intermediate steps or reliably integrate external knowledge, leading to propagated errors and misleading conclusions. Inspired by Bayesian inference logic in clinical diagnosis, we propose **BELIEF**, a **B**ayesian **E**vidence-**L**everaged **I**nfErence **F**ramework that enables LLMs to perform reliable medical reasoning through dynamic belief assessment and knowledge-augmented disambiguation. First, the Bayesian Belief Updating Module (BBUM) treats diagnostic hypotheses as prior probabilities and calculates posterior probabilities based on observed clinical evidence, preventing early-stage error propagation with statistical rigor. Second, the Dynamic Evidence Retrieval Module (DERM) autonomously queries external medical databases to resolve ambiguity when the posterior belief exhibits high entropy. Experiments across 7 datasets and 2 LLM backbones show that BELIEF significantly improves diagnostic accuracy, enhances interpretability, and mitigates hallucinations through explicit probabilistic updates.

1 Introduction

Medical diagnosis is a fundamental aspect of healthcare, in which accurate clinical reasoning is essential for patient health and survival (Tang et al., 2024). Recently, Large Language Models (LLMs) have demonstrated remarkable proficiency in the medical domain, often matching or exceeding human-level performance on standardized medical licensing examinations (Chen et al., 2023; Dou et al., 2025; Toma et al., 2023). They also offer a promising pathway to extend high-quality medical knowledge and decision support to underserved

and medically underdeveloped regions (Zhou et al., 2024).

Despite their impressive linguistic capabilities, a critical bottleneck persists: these models are inherently prone to hallucinations during complex, multi-step clinical reasoning. In clinical domains where errors may result in loss of human life, the lack of transparency in LLM reasoning, coupled with a tendency toward overconfidence in incorrect reasoning trajectories, continues to hinder their adoption (Griot et al., 2025; Johri et al., 2025).

Current research in medical LLMs primarily focuses on Chain-of-Thought (CoT) (Sun et al., 2025; Liu et al., 2024; Liu et al.) and Retrieval-Augmented Generation (RAG) (Wu et al., 2025a; Ke et al., 2025). While these methods improve performance, they suffer from two fundamental weaknesses. First, the absence of intermediate confidence assessment: existing models typically proceed through a reasoning chain in a linear manner, but do not provide explicit estimates of uncertainty for intermediate steps. This causes the reasoning process to follow an erroneous trajectory, with a single early-stage error propagating and intensifying over time. Second, the lack of strategic knowledge integration: models often struggle to recognize their own epistemic boundaries, either failing to retrieve external knowledge when faced with ambiguity or indiscriminately incorporating irrelevant data that can mislead the final diagnosis (Hu et al., 2025; Niu et al., 2024).

To overcome these barriers, medical AI must transition from simple sequence generation to a deliberative slow-thinking paradigm that emulates the cognitive rigor of expert physicians. In clinical practice, a human physician does not merely follow a fixed script; instead, they employ a Bayesian-like diagnostic workflow (Kassirer, 1989; Bours, 2021; Gill et al., 2005). They start with a prior hypothesis based on initial symptoms and iteratively update their belief as new clinical evidence

*Equal Contributions.

†Corresponding author.

(labs, medical guidelines, history records) emerges. When the internal confidence falls below a certain threshold, the physician consults authoritative medical guidelines to resolve ambiguity. Inspired by this systematic reasoning process, we propose **BELIEF** (**B**ayesian **E**vidence-**L**everaged **I**nfErence **F**ramework), a novel framework designed to instill statistical rigor and dynamic self-correction into LLM-based medical reasoning.

The BELIEF framework introduces two core innovations: (1) Bayesian Belief Updating Module (BBUM): Unlike traditional deterministic reasoning, BBUM treats each diagnostic step as a probabilistic inference task. By calculating the posterior probability of a hypothesis based on observed clinical evidence, the model can quantitatively assess its diagnostic belief, ensuring that the reasoning path is grounded in statistical consistency. (2) Dynamic Evidence Retrieval Module (DERM): This module grants the model the autonomy to recognize uncertainty. When the belief score derived from BBUM indicates high ambiguity, DERM triggers targeted queries to external, authoritative medical databases (e.g., history records, clinical guidelines), integrating real-time evidence to refine the reasoning process.

We evaluate the **BELIEF** framework across 7 diverse medical diagnosis datasets and 2 LLM backbones. Our experimental results demonstrate that BELIEF significantly outperforms existing strong baselines in diagnostic accuracy. More importantly, we show that by explicitly modeling the evolution of beliefs and providing evidence traceability, our framework substantially mitigates hallucinations and enhances the interpretability of AI-driven diagnoses. To sum up, our contributions can be summarized as follows:

- We propose a deliberative reasoning architecture that models medical diagnosis as a continuous probabilistic refinement process.
- We develop a novel framework that establishes a closed-loop mechanism for diagnostic inference. By leveraging Bayesian belief to guide the acquisition of evidence, which in turn feeds back to modulate and sharpen belief distributions toward final diagnostic certainty.
- Our approach not only improves diagnostic accuracy but also establishes more robust inference chains, enhances interpretability, and effectively mitigates hallucinations.

2 Related work

Trustworthy Medical LLM Reasoning LLMs in clinical settings are prone to hallucinations caused by pretraining errors and overconfident reward mechanisms, with medical risks amplified by data scarcity, demographic bias, and underspecified prompts (Kalai et al., 2025; Zhu et al., 2025; Zhang et al., 2024; Chen et al., 2025). Existing mitigation strategies primarily focus on post-hoc interventions, such as semantic inconsistency detection (Farquhar et al., 2024), ensemble-based voting (Till et al., 2025), and RAG grounded in authoritative corpora (Zhang et al., 2024; Xia et al., 2025). Recent work therefore turns to trajectory-level supervision and human-centric alignment, using evidence graphs, tree-search (Mu et al., 2025; Jiang et al., 2026), verifier-guided learning, multi-objective rewards (Chen et al., 2024a; Gu et al., 2025), physician-annotated trajectories, reflective prompting, and multi-agent collaboration (Wang et al., 2024; Zhao et al., 2024; Tang et al., 2024). However, current Slow-Thinking Med-LLMs still lack quantifiable uncertainty assessment and confidence-aware triggers, unlike clinicians who deliberate when ambiguity or conflicting evidence is detected. Medical diagnosis can be viewed as a Bayesian inference process (Gill et al., 2005), where beliefs are updated as new evidence arrives. This motivates our approach. Prior work (Guez et al., 2012; Lee et al., 2018; Zhang et al., 2026) studied BAMDPs; here, we apply Bayesian belief updating to medical diagnosis.

Medical Retrieval-Augmented Generation Systems Recent frameworks have introduced multi-agent iterative retrieval, triple-graph constructions, and hierarchical community-level search to bridge the gap between private clinical data and external medical corpora (Wang et al., 2025; Wu et al., 2025a; Jiang et al., 2025). However, retrieval quality is critical: insufficient retrieval may induce hallucinations (Rashtchian and Juan, 2025), whereas excessive retrieval can introduce noisy or misleading evidence (Amiraz et al., 2025). Thus, recent Adaptive RAG and Self-Reflective methods optimize retrieval policies by estimating information sufficiency and internal knowledge gaps (Yang et al., 2025b; Tan et al., 2024; Zheng et al., 2025). Building on this line of work, BELIEF formulates medical diagnosis as continuous probabilistic refinement, dynamically calibrating evidence retrieval according to real-time uncertainty.

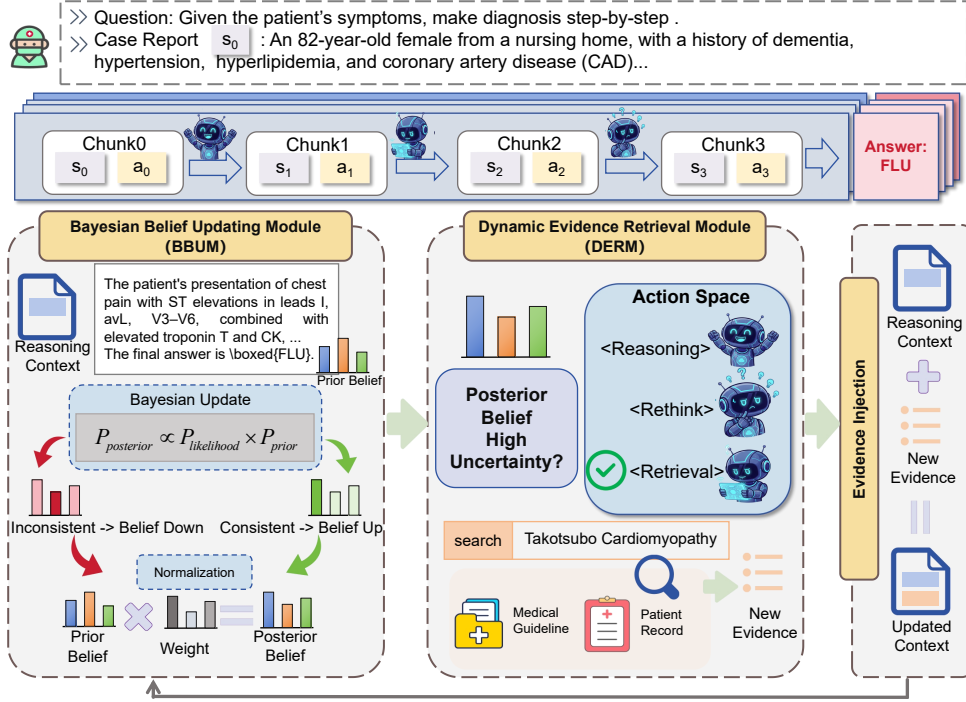


Figure 1: Overview of BELIEF, which mainly consists of 2 modules: (1) Bayesian Belief Updating Module (BBUM): dynamically quantifies diagnostic uncertainty and leverages these assessments to strategically trigger external knowledge retrieval; (2) Dynamic Evidence Retrieval Module (DERM): activates during high-uncertainty states to incorporate external knowledge into the reasoning context, thereby reducing diagnostic uncertainty through evidence-based belief refinement. The BBUM and DERM form a unified diagnostic framework where the former triggers the latter upon detecting high uncertainty, while the latter provides the external evidence necessary for the former to refine its beliefs.

3 Framework

3.1 Framework Overview

BELIEF leverages Bayesian belief updating to provide training-time supervision for medical diagnostic reasoning. As shown in Figure 1, BBUM models the posterior distribution over candidate diagnoses and derives uncertainty signals to guide both reasoning and evidence retrieval through a unified reward. DERM further learns when to trigger actions and how to incorporate retrieved evidence into subsequent reasoning. Together, these components form a closed-loop training framework that couples diagnostic reasoning with selective evidence acquisition.

3.2 Task Formulation

We formulate medical diagnosis with LLMs as a sequential probabilistic inference problem over a set of candidate diagnoses $\mathcal{Y} = \{y_1, \dots, y_K\}$. Given an initial patient record X_0 , the goal is to identify the most plausible diagnosis by progressively incorporating clinical evidence. At each step t , the model maintains a context $C_t = \{X_0, r_{1:t}, a_{1:t}, e_{1:t}\}$, in-

cluding the patient information, reasoning traces, actions, and accumulated evidence. Conditioned on C_t , the LLM performs reasoning and selects an action to acquire new evidence. From a probabilistic perspective, the model maintains a belief over candidate diagnoses, initialized by $P(y_k | X_0)$ and iteratively refined as new evidence is observed. The final prediction corresponds to the diagnosis with the highest posterior belief under the full context.

3.3 Bayesian Belief Updating Module

Candidate Diagnosis Pool Construction. During training, given a patient case X_0 , the LLM generates a candidate diagnosis set $\mathcal{Y}(X_0) = y_1, \dots, y_K$. If the ground-truth diagnosis y^* is not included in the set, we append it, as BBUM requires an explicit target to provide a valid supervision signal. Each hypothesis y_k is initialized with a prior belief $b_0(y_k) = P(y_k | X_0)$. All candidate diagnoses share the same b_0 .

Chunk-wise evidence assessment. We partition the reasoning trajectory C into T non-overlapping chunks of fixed size $\Delta = 128$ ($T = \lceil |C|/\Delta \rceil$). At

each chunk t , conditioned on the current context C_t , the LLM performs reasoning and selects an action a_t (e.g., finding similar cases, or lookup knowledge), which yields new clinical evidence e_{t+1} .

For each candidate y_k , the step-wise evidence likelihood $\ell_k(t) = \frac{p_k(t)}{p_k(t-1)}$, where $p_k(t) = P_\theta(y_k | C_t)$ is the LLM’s diagnostic confidence in hypothesis y_k after reading the accumulated clinical context C_t . $\ell_k(t)$ measures how likely the newly acquired evidence would be observed under hypothesis y_k .

We maintain an unnormalized belief weight $\tilde{b}_t(y_k)$ for each diagnosis, which is recursively updated by accumulating evidence likelihoods: $\tilde{b}_t(y_k) = \tilde{b}_{t-1}(y_k) \cdot \ell_k(t)$. This recursion admits a direct Bayesian interpretation, where each newly acquired piece of evidence contributes multiplicatively to the posterior belief. The normalized belief is then obtained as $b_t(y_k) = \frac{\tilde{b}_t(y_k)}{\sum_j \tilde{b}_t(y_j)}$, which corresponds to the posterior probability $P(y_k | C_t)$ over candidate diagnoses. After multiple reasoning and evidence acquisition steps, the final belief is given by $b_T(y_k) \propto P(y_k | X_0) \prod_{t=1}^T \ell_k(t)$, which integrates prior knowledge with accumulated clinical evidence. The final diagnosis is selected as $\hat{y} = \arg \max_{y_k} b_T(y_k)$. This formulation models diagnosis as a sequential process of hypothesis generation and evidence-driven belief revision, where each step incrementally refines the posterior distribution over candidate diseases.

Uncertainty signal and posterior advantage.

We elevate the Shannon entropy (Shannon, 1948) of the diagnostic posterior $H_t = -\sum_{k=1}^K b_t(c_k) \log b_t(c_k)$, to an externally readable signal. BELIEF treats H_t as the canonical uncertainty interface between reasoning and action: H_t near 0 indicates the model has converged on a confident differential and should commit, whereas H_t near $\log K$ indicates the chain is exploring without resolution and external evidence should be sought. By making H_t explicit, the retrieval policy in DERM becomes a *learnable* behavior governed by Bayesian uncertainty, rather than a hand-crafted prompt-based heuristic.

Let T denote the index of the final chunk. We define $A_t = \sum_{k=1}^K (b_T(y_k) - b_t(y_k)) * b_t(C_t)$,

BBUM therefore yields two complementary signals: (i) the chunk-level posterior advantage A_t , which is assigned to all tokens within chunk; (ii)

the entropy signal H_t . These two quantities are subsequently used by the unified objective in section 3.5 and the retrieval policy in section 3.4, respectively.

3.4 Dynamic Evidence Retrieval Module

DERM turns the BBUM uncertainty signal H_t into a learnable retrieval action. Unlike prior approaches where retrieval is triggered either by token-level heuristics (e.g., Self-RAG (Asai et al., 2024)), uncertainty signals (e.g., FLARE (Jiang et al., 2023)), or intermediate reasoning steps (e.g., Search-R1 (Jin et al., 2025), R1-Searcher (Song et al., 2025)), and where supervision is typically applied at the token or trajectory level, DERM derives both its retrieval trigger and reward from an explicit Bayesian posterior over diagnoses. This allows retrieval decisions to be directly aligned with the diagnostic objective, and enables joint optimization of retrieval and diagnosis under a shared advantage signal.

Retrieval as an environment-controlled action.

To realize uncertainty-triggered retrieval as a learnable behavior, we extend the action space $\mathcal{A} = \mathcal{V}$ with two reserved tokens $\langle \text{search} \rangle$ and $\langle / \text{search} \rangle$, which are newly embedded and incorporated into the LM head via resizing. Under the uncertainty gated reward, the policy is encouraged to emit this token pair only when its posterior entropy H_t surpasses a predefined threshold η . Mechanically, whenever $\langle \text{search} \rangle, q, \langle / \text{search} \rangle$ is emitted, generation pauses, the query q is sent to a medical RAG backend $\mathcal{R}$, and the top- k passages $e_t = \mathcal{R}(q)$ are spliced into context inside an environment-controlled $\langle \text{results} \rangle, e_t, \langle / \text{results} \rangle$ block. The action mask is zeroed on observation tokens so retrieved text acts as an exogenous state transition rather than a behavior to imitate.

Uncertainty-gated reward and closed-loop refresh.

To avoid over-querying, triggering and reward are fused into one signal from H_t . Let t^-, t^+ be the chunks immediately before the search and after ingestion:

$$r_t^{\text{search}} = \lambda [\mathbb{1}_{H_{t^-} > \eta} (H_{t^-} - H_{t^+}) - \beta \mathbb{1}_{H_{t^-} \leq \eta}] ,$$

with threshold η , over-querying penalty β , and a budget of M searches per trajectory. The positive branch rewards searches that are *both* justified ($H_{t^-} > \eta$) and effective ($H_{t^-} - H_{t^+} > 0$). Following

ingestion, BBUM recomputes $p_k(t^+)$ on the updated state $s_{t+} = s_{t-} \oplus e_t$, producing a sharpened posterior b_{t+} and an increased calibration advantage A_{t+} over downstream tokens. This sharpening concurrently reduces entropy H_{t+} and eliminates the incentive for redundant follow-up queries. Triggering, reward, and self-suppression thus arise as different manifestations of a single Bayesian posterior, closing the loop between BBUM and DERM without introducing an auxiliary reward model.

3.5 Unified Training Objective

We optimize π_θ with a clipped policy-gradient objective whose advantage aggregates four reward streams aligned with the three actions of our MDP: (i) BBUM’s chunk-level posterior advantage A_t , which credits the *reasoning* action; (ii) DERM’s uncertainty-gated retrieval reward r_s^{search} for the *retrieval* action; (iii) a parallel *rethink* reward $r_s^{\text{rethink}} = \delta \mathbb{1}[\Delta p_{y^*}(t_s^+) > \Delta p_{y^*}(t_s^-)]$ that fires only when a <wait> action accelerates per-chunk convergence to the ground truth (with Δp_{y^*} , so that self-reflection is incentivized only when it is empirically productive; and (iv) a trajectory-level outcome shaping term R_{out} returned by the medical verifier, jointly scoring answer format and embedding semantic correctness. We aggregate these into a single per-token advantage,

$$A_t^{\text{tot}} = A_t + R_{\text{out}}(\tau) + \sum_{s \in \mathcal{S}_t^{\text{search}}} r_s^{\text{search}} + \sum_{s \in \mathcal{S}_t^{\text{rethink}}} r_s^{\text{rethink}},$$

where $\mathcal{S}_t^{\text{search}}, \mathcal{S}_t^{\text{rethink}}$ index the retrieval and rethink actions whose effects reach chunk t , and the floor index broadcasts A to every token of its chunk. The clipped policy update

$$\mathcal{L}_\theta = -\mathbb{E}_t[\min(\rho_t A_t^{\text{tot}}, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) A_t^{\text{tot}})],$$

with $\rho_t = \pi_\theta(a_t | s_t) / \pi_{\theta_{\text{old}}}(a_t | s_t)$, jointly trains all three actions from one gradient signal, regularized by a KL penalty against a frozen reference policy. Because the three Bayesian streams (A , r_s^{search} , r_s^{rethink}) all derive from the same posterior b_t and its per-chunk dynamics, BELIEF reduces RL fine-tuning of a clinical reasoner to a single actor optimization in which reasoning, reflection, and evidence acquisition are end-to-end optimized under one interpretable Bayesian advantage. After optimization, inference proceeds through the learned policy π_θ , which directly generates reasoning trajectories, <search> actions and queries, and final diagnoses. The candidate set, b_t , and H_t are used only to provide training-time supervision.

4 Experiments

4.1 Benchmarks

We performed evaluations on a suite of established medical diagnosis benchmarks structured into two primary domains: (1) in-domain benchmarks (ID): MIMIC (Johnson et al., 2023), PMC-Patients (Zhao et al., 2022), RareArena (Chen et al., 2026), RareBench (Chen et al., 2024b); (2) out-of-domain benchmarks (OOD): DiagArena (Zhu et al., 2026), MedCase (Wu et al., 2025b). Following Zheng et al. (2025), we categorized MIMIC into common diseases (MIMIC-C) and rare diseases (MIMIC-R). MIMIC-C and PMC-Patients are used as benchmarks for common diseases, while MIMIC-R, RareArena, and RareBench serve as benchmarks for rare diseases.

4.2 Experimental Setup

4.2.1 Baselines

We choose the following 4 categories to serve as baselines: (1) Large language models (>10B): GPT-5.4, GPT-4o-mini (Achiam et al., 2023), GPT-3.5-Turbo (Ouyang et al., 2022), DeepSeek-R1-0528 (DeepSeek-AI, 2025), Qwen-3-32B (Yang et al., 2025a); (2) Small language models (<10B): Qwen2.5-7B (Team, 2024), Qwen2.5-7B w/ RAG, Llama3.1-8B (Grattafiori et al., 2024), Llama3.1-8B w/ RAG, Qwen3.5-9B (Team, 2026), Self-RAG-Llama-7B (Asai et al., 2024); (3) Medical Instruct Models(<10B): UltraMedical3-8B (Zhang et al., 2024), OpenBioLLM-8B (Pal and Sankarabsubbu, 2024), Med42-8B (Christophe et al., 2024); (4) Medical Reasoning Models(<10B): Huatuo-o1-8B (Chen et al., 2024a), Med-s³ (Jiang et al., 2026), Med-CEG (Mu et al., 2025), Deep-DxSearch-8B (Zheng et al., 2025). The implementation details and experimental settings are provided in Appendix C.

4.2.2 Evaluation Metrics

Following Zheng et al. (2025); Shikino et al. (2024); Bhasuran et al. (2025), we adopt Acc@1 and Acc@5 as the main evaluation metrics to assess the performance of our proposed method. Acc@K measures whether the ground-truth label appears within the top-K predicted results generated by the model. Separately from diagnostic benchmarking, following Mu et al. (2025), we assess reasoning reliability using 4 independent LLM judges: GPT-5.5, GPT-3.5-Turbo (Ouyang et al., 2022), DeepSeek-R1-0528 (DeepSeek-AI, 2025),

Models	Training Type	MIMIC-C		PMC-Patients		MIMIC-R		RareArena		RareBench	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
[rgb] .851, .851, .851	Large language models (>10B)										
GPT-5.4	-	21.80	60.44	17.60	62.92	26.70	66.39	31.00	89.80	24.07	67.28
GPT-4o-mini	-	30.58	74.77	16.93	58.67	30.93	71.13	20.30	60.90	15.53	62.11
GPT-3.5-Turbo	-	25.83	64.98	18.40	59.07	31.86	70.00	27.00	77.40	16.77	72.05
DeepSeek-R1-0528	-	14.63	36.13	12.40	25.33	26.19	45.36	20.40	61.40	16.15	40.99
Qwen-3-32B	-	31.18	69.73	14.00	58.13	29.79	67.11	18.00	62.30	19.88	72.67
[rgb] .851, .851, .851	Small language models (<10B)										
Qwen2.5-7B	-	4.23	8.61	10.96	24.45	9.63	18.77	6.06	12.80	17.86	30.48
Qwen2.5-7B w/ RAG	-	9.13	12.62	20.08	32.25	17.82	25.17	8.28	13.02	31.98	57.70
Qwen3.5-9B	-	7.87	8.27	12.73	13.00	9.27	9.70	8.40	8.80	13.50	14.19
Llama3.1-8B	-	3.81	8.00	15.13	25.08	3.40	12.75	7.94	12.87	20.89	35.71
Llama3.1-8B w/ RAG	-	9.01	13.03	25.77	40.55	15.03	22.10	11.31	17.02	33.27	55.01
Self-RAG-Llama-7B	-	30.54	32.66	28.61	31.12	32.54	35.03	10.78	12.41	39.51	54.32
[rgb] .851, .851, .851	Medical Instruct Models(<10B)										
UltraMedical3-8B	SFT+DPO+KTO	20.79	38.85	14.67	33.33	23.92	41.13	25.90	38.10	30.43	37.89
OpenBioLLM-8B	-	18.16	34.51	11.20	24.00	15.57	29.38	14.80	24.10	18.63	30.43
Med42-8B	SFT+DPO	20.48	39.35	18.53	32.93	25.26	33.92	18.30	26.00	22.36	32.92
[rgb] .851, .851, .851	Medical Reasoning Models(<10B)										
Huatuo-o1-8B	SFT+PPO	18.26	38.14	18.67	39.87	20.10	41.86	24.00	40.20	21.12	47.83
Med-s ³	SFT+DPO	21.70	50.35	26.27	<u>42.53</u>	18.45	40.10	<u>30.60</u>	<u>44.70</u>	31.68	45.96
Med-CEG	SFT+GRPO	16.75	42.08	16.67	22.93	23.40	43.40	17.60	35.00	34.16	39.13
Deep-DxSearch-8B	SFT+GRPO	21.09	27.85	34.13	45.73	41.96	55.05	22.40	30.00	64.60	73.91
BELIEF-Qwen2.5-7B	SFT+PPO	<u>54.29</u>	<u>65.79</u>	<u>29.47</u>	42.13	60.21	<u>69.90</u>	29.40	44.40	55.90	<u>73.29</u>
BELIEF-Llama3.1-8B	SFT+PPO	58.32	69.42	28.40	40.80	<u>60.00</u>	70.41	34.80	49.50	<u>59.01</u>	72.67

Table 1: **The overall performance of baselines and BELIEF on 5 in-domain benchmarks.** We highlight the best results with **bold** and underline the second best results among models with a similar size. Our proposed method achieves remarkable performance across 3 benchmarks and yields competitive, second best results on the remaining 2 benchmarks, demonstrating robust effectiveness compared to existing baselines.

Models	Diagnosis Arena		Medcase	
	Acc@1	Acc@5	Acc@1	Acc@5
Huatuo-o1-8B	9.62	32.79	9.12	<u>33.76</u>
Med-s ³	10.39	32.35	9.88	27.37
Med-CEG	10.93	26.31	7.47	20.11
Deep-DxSearch-8B	9.07	33.99	6.20	27.70
BELIEF-Qwen2.5-7B	15.41	35.83	<u>13.87</u>	32.72
BELIEF-Llama3.1-8B	<u>14.64</u>	<u>35.15</u>	16.79	34.82

Table 2: **The overall performance of Medical Reasoning Models and BELIEF on 2 out-of-domain benchmarks.**

and Qwen-2.5-72B (Bai et al., 2023). Each response is rigorously scored across five dimensions: (i) Logical Coherence, (ii) Factual Accuracy, (iii) Evidence Faithfulness, (iv) Interpretability & Clarity, and (v) Information Utilization. Prompt template is available in Table 6 (Appendix B).

4.3 Main Results

4.3.1 Evaluation of Diagnosis Accuracy

We conduct comprehensive experiments on five in-domain and two out-of-domain benchmarks to evaluate BELIEF for medical diagnosis, with results reported in Tables 1 and 2. Overall, Medical Reasoning Models consistently outperform SLMs and Medical Instruct Models at comparable scales, highlighting the importance of specialized reasoning mechanisms in the medical domain. On challenging benchmarks such as MIMIC-C and

RareBench, our models achieve performance close to, or even surpassing, much larger LLMs. These results highlight the effectiveness of BELIEF in improving the medical reasoning capability of smaller LLMs. Moreover, BELIEF is model-agnostic and can be readily extended to stronger trainable open-source backbones.

Self-RAG-Llama-7B, as an autonomously retrieving model, achieves strong performance on Acc@1; however, its gains on Acc@5 are relatively limited, possibly due to insufficient diversity in retrieved candidates. In contrast, although Qwen3.5-9B is a more recent model, its inferior performance compared to Qwen2.5-7B may be attributed to overly long reasoning chains. Notably, on MIMIC-R, ours-Llama3.1-8B attains the best Acc@1 of 60.21% among 8B models, indicating improved handling of complex clinical logic beyond factual memorization. In OOD settings, our models also show strong Acc@1 performance, demonstrating precise and generalizable diagnostic reasoning across unseen clinical distributions. These results confirm the robustness and accuracy of BELIEF, suggesting that BBUM effectively stabilizes reasoning chains and mitigates the impact of noisy clinical signals.

Beyond accuracy, we further evaluate whether the model reduces clinically risky hallucinations (The description of the evaluation parameters is

Models	MIMIC-C		PMC-Patients		MIMIC-R		RareArena		RareBench	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
SFT Policy	54.25	66.80	26.54	38.95	59.81	67.15	33.20	48.30	55.48	69.20
w/o BBUM	54.09	66.90	26.67	39.07	58.97	67.11	32.60	48.10	51.55	68.32
w/o DERM	56.33	69.05	26.91	40.16	59.87	67.64	34.62	48.65	56.21	70.81
BELIEF-Llama3.1-8B	58.32	69.42	28.40	40.80	60.00	70.41	34.80	49.50	59.01	72.67

Table 3: Ablation study for BBUM and DERM, highlighting the contribution of each component.

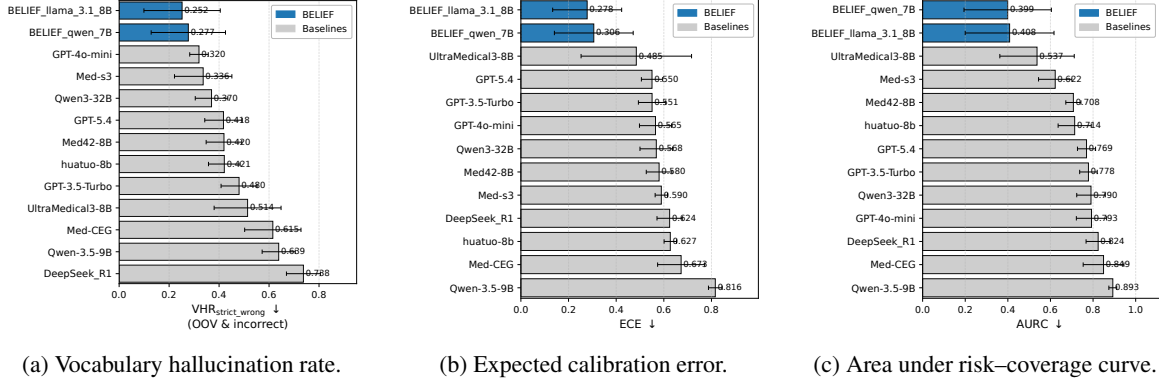


Figure 2: Cross-model hallucination and calibration comparison on five clinical-diagnosis benchmarks (means over datasets, error bars = std). Lower is better for all three metrics.

provided in Appendix E). Figure 2 shows that our models achieve top-tier performance, with reductions of 21%, 43%, and 24% in hallucinations, calibration error, and selective risk, respectively, suggesting improvements beyond prediction accuracy in reliability and uncertainty awareness.

Overall, our proposed BELIEF shows strong and consistent performance on most benchmarks, highlighting its robustness and accuracy. This indicates that the BBUM effectively stabilizes the reasoning chain, preventing the model from being led astray by noisy clinical signals.

4.3.2 Qualitative Analysis of Diagnosis Reasoning

Figure 3 presents a comparative analysis using 4 LLMs as evaluators. Across all six baselines, although absolute scores vary across different LLM evaluators, the overall performance trend remains consistent. Our proposed BELIEF framework outperform models of comparable scale, demonstrating that our optimization methods significantly enhance reasoning capabilities. Furthermore, multi-dimensional evaluation via GPT-3.5-Turbo (Figure 5) confirms that our method consistently outperforms baselines, particularly in **Interpretability & Clarity** and **Evidence Faithfulness**. These gains indicate that our approach produces more transparent, well-structured, and evidence-grounded reasoning paths, resulting in more reliable and logi-

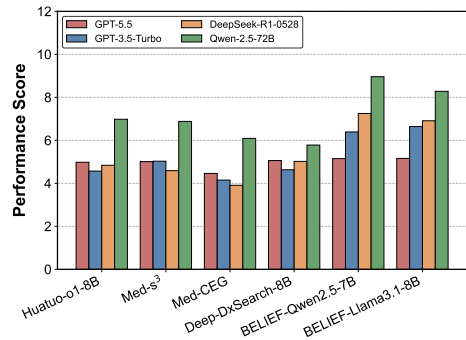


Figure 3: LLM-based evaluation of the reasoning process for different models.

cally coherent diagnostic responses. We evaluated human-LLM agreement on 100 randomly sampled cases. Five experts independently scored them using the same criteria, and alignment was measured with Spearman’s ρ . The LLM-expert correlation is 0.77, while the average inter-expert ρ is 0.80, indicating that the LLM is a reliable proxy for human judgment.

4.4 Ablation Study

To verify the effectiveness of BELIEF’s components, we conduct an ablation study with the following variants: (1) SFT Policy: using only the SFT policy; (2) w/o BBUM: incorporating retrieval mechanisms into the Cold Start SFT policy; and (3) w/o DERM: performing inference based solely

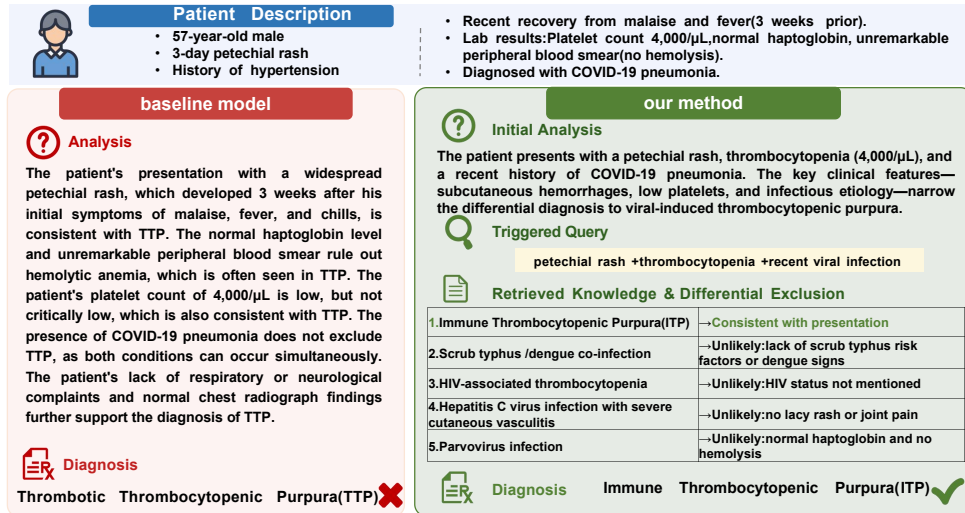


Figure 4: **Case study of clinical diagnosis.** The baseline model confidently predicts the incorrect diagnosis TTP, while BELIEF triggers evidence retrieval, rules out inconsistent candidates, and correctly identifies ITP. The retrieval query is highlighted in yellow.

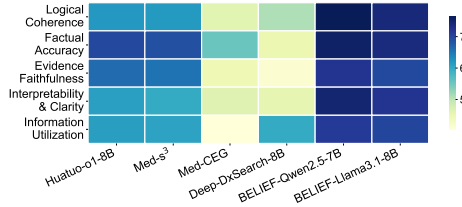


Figure 5: **Comparative analysis of reasoning quality across 5 key dimensions.** Darker colors indicate better performance. Our proposed method also exhibits superior proficiency in terms of reasoning quality.

on the BBUM without external retrieval.

We summarize the ablation results in Table 3. Removing BBUM leads to the weakest performance, highlighting the importance of dynamically updating posterior beliefs. In contrast, using reasoning alone suffers from information insufficiency. By integrating reasoning and retrieval into a unified cognitive flow, BELIEF achieves the best overall performance among all variants. A paired t-test further confirms that the improvement is statistically significant ($t = -6.0963$, $p < 0.001$), indicating that the gains are reliable. We additionally analyze sensitivity to the entropy threshold η in Appendix D.

4.5 Case Study

In Figure 4, we present a diagnostic case, with full details provided in Table 5 (Appendix A). The initial patient description suggests an abnormal platelet count, but the evidence is insufficient for a confident diagnosis. This uncertainty triggers ev-

idence retrieval. Based on the retrieved evidence, BELIEF systematically rules out inconsistent diseases and correctly diagnoses Immune Thrombocytopenic Purpura (ITP). In contrast, without our method, the model consistently predicts Thrombotic Thrombocytopenic Purpura (TTP) across 5 trials. Although ITP and TTP differ by only one word, this error can lead to inappropriate treatment and serious clinical risks. This case shows that repeated sampling cannot easily eliminate such confident hallucinations, further demonstrating the necessity of our approach.

5 Conclusion

This paper introduced BELIEF, a reasoning framework that models medical diagnosis as a continuous trajectory of probabilistic refinement. At its core, the framework establishes a recursive feedback loop between internal diagnostic hypothesis generation and external evidence verification. The Bayesian Belief Updating Module (BBUM) serves as the cognitive engine that quantifies uncertainty, which then parametrically triggers the Dynamic Evidence Retrieval Module (DERM) to resolve specific clinical ambiguities. Experimental results across 7 diverse medical benchmarks and 2 backbones demonstrate that this interlocked architecture significantly surpasses existing strong baselines. Beyond accuracy, BELIEF provides a mathematically grounded trace of diagnostic confidence, reducing hallucinations.

Limitations

While our method achieves promising results, several limitations remain: (1) Our approach is strictly confined to textual medical data. It is currently incapable of processing complex, unstructured, and multimodal clinical inputs (e.g., medical imaging or waveforms). (2) Although BELIEF is not tied to a specific backbone and can in principle be instantiated with stronger trainable open-source LLMs, our empirical validation is currently limited to public datasets and two 7–8B backbones. Its scalability and effectiveness on substantially larger models (e.g., 70B+) remain to be verified.

Ethical considerations

We performed evaluations on a suite of established medical diagnosis benchmarks, including MIMIC, PMC-Patients, RareArena, RareBench, Diagnosis Arena and MedCase. This research adheres to strict ethical standards in dataset usage and evaluation. The MIMIC dataset is utilized under its data use agreement, ensuring compliance with all relevant regulations and responsible handling of de-identified clinical data. Other datasets are derived from publicly available biomedical literature or prior benchmark studies, and are used in accordance with their respective terms of use.

This work is intended for research purposes only and does not constitute medical advice. The proposed model is not validated for clinical use and should not be used for diagnosis or treatment decisions.

Acknowledgments

We sincerely thank all collaborators for their valuable support and insightful discussions.

References

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

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karni. 2025. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18228–18258.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to re-

trieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Abhinand Balachandran. 2024. *Medembed: Medical-focused embedding models*.

Balu Bhasuran, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu, and Zhe He. 2025. Preliminary analysis of the impact of lab results on large language model generated differential diagnoses. *npj Digital Medicine*, 8(1):166.

Martijn JL Bours. 2021. Bayes’ rule in diagnosis. *Journal of Clinical Epidemiology*, 131:158–160.

Jaime Carbonell and Jade Goldstein. 1998. *The use of MMR, diversity-based reranking for reordering documents and producing summaries*. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’98*, page 335–336, New York, NY, USA. Association for Computing Machinery.

Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, Honghan Wu, Bin Sheng, Shuyang Zhang, Sheng Yu, and Tien Yin Wong. 2026. *Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis*. *The Lancet Digital Health*, page 100953.

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

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

Xiuying Chen, Tairan Wang, Juexiao Zhou, Zirui Song, Xin Gao, and Xiangliang Zhang. 2025. Evaluating and mitigating bias in AI-based medical text generation. *Nature Computational Science*, pages 1–9.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024b. Rarebench: Can LLMs serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4850–4861.

- Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, and 1 others. 2024. Med42—evaluating fine-tuning strategies for medical LLMs: Full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Christopher J Gill, Lora Sabin, and Christopher H Schmid. 2005. Why clinicians are natural bayesians. *Bmj*, 330(7499):1080–1083.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.
- Boyang Gu, Hongjian Zhou, Bradley Max Segal, Jinge Wu, Zeyu Cao, Hantao Zhong, Lei Clifton, Fenglin Liu, and David A Clifton. 2025. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. *arXiv preprint arXiv:2512.00601*.
- Arthur Guez, David Silver, and Peter Dayan. 2012. Efficient bayes-adaptive reinforcement learning using sample-based search. *Advances in Neural Information Processing Systems*, 25.
- Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853.
- Pengcheng Jiang, Cao Danica Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2025. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations*, volume 2025, pages 81785–81830.
- Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2026. Meds³: Towards medical slow thinking with self-evolved soft dual-sided process supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31319–31327.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.
- Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86.
- Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.
- Jerome P Kassirer. 1989. Diagnostic reasoning. *Annals of Internal Medicine*, 110(11):893–900.
- Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, and 1 others. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.
- Gilwoo Lee, Brian Hou, Aditya Mandalika, Jeongseok Lee, Sanjiban Choudhury, and Siddhartha S Srinivasa. 2018. Bayesian policy optimization for model uncertainty. *arXiv preprint arXiv:1810.01014*.
- Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.
- Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, SU Yihang, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael Lyu, and Wenting Chen. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. In *The Thirtieth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

- Linjie Mu, Yannian Gu, Zhongzhen Huang, Yakun Zhu, Shaoting Zhang, and Xiaofan Zhang. 2025. Medceg: Reinforcing verifiable medical reasoning with critical evidence graph. *arXiv preprint arXiv:2512.13510*.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences.
- Cyrus Rashtchian and Da-Cheng Juan. 2025. Deeper insights into retrieval-augmented generation: The role of sufficient context. *Google Research Blog*.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Kiyoshi Shikino, Taro Shimizu, Yuki Otsuka, Masaki Tago, Hiromizu Takahashi, Takashi Watari, Yosuke Sasaki, Gemmei Iizuka, Hiroki Tamura, Koichi Nakashima, and 1 others. 2024. Evaluation of ChatGPT-generated differential diagnosis for common diseases with atypical presentation: Descriptive research. *JMIR Medical Education*, 10:e58758.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.
- Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and 1 others. 2025. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26457–26478.
- Jiejun Tan, Zhicheng Dou, Yutao Zhu, Peidong Guo, Kun Fang, and Ji-Rong Wen. 2024. Small models, big insights: Leveraging slim proxy models to decide when and what to retrieve for llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4420–4436.
- 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. 2024. *Qwen2.5: A party of foundation models*.
- Qwen Team. 2026. *Qwen3.5: Accelerating productivity with native multimodal agents*.
- Demian Till, John Smeaton, Peter Haubrick, Gouse Saheb, Florian Graef, and David Berman. 2025. Teaming llms to detect and mitigate hallucinations. *arXiv preprint arXiv:2510.19507*.
- Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.
- Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. *Advances in Neural Information Processing Systems*, 37:74999–75011.
- Sheng Wang, Fangyuan Zhao, Dechao Bu, Yunwei Lu, Ming Gong, Hongjie Liu, Zhaohui Yang, Xiaoxi Zeng, Zhiyuan Yuan, Baoping Wan, and 1 others. 2025. Lins: A general medical q&a framework for enhancing the quality and credibility of llm-generated responses. *Nature communications*, 16(1):9076.
- Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025a. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.
- Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025b. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Y Zou, and Huaxiu Yao. 2025. Mmed-rag: Versatile multimodal rag system for medical vision language models. In *International Conference on Learning Representations*, volume 2025, pages 66188–66217.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

- Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. 2025b. Knowing you don't know: Learning when to continue search in multi-round rag through self-practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1305–1315.
- Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.
- Shenao Zhang, Yaqing Wang, Yinxiao Liu, Tianqi Liu, Peter Grabowski, Eugene Ie, Zhaoran Wang, and Yunxuan Li. 2026. Beyond markovian: Reflective exploration via bayes-adaptive rl for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pages 120275–120294.
- Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-reflection (far) improves confidence calibration of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8702–8718.
- Zhengyun Zhao, Qiao Jin, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient notes and relations extracted from case reports in pubmed central. *arXiv preprint arXiv:2202.13876*.
- Qiaoyu Zheng, Yuze Sun, Chaoyi Wu, Weike Zhao, Pengcheng Qiu, Yongguo Yu, Kun Sun, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2025. End-to-end agentic rag system training for traceable diagnostic reasoning. *arXiv preprint arXiv:2508.15746*.
- Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and 1 others. 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649.
- Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2026. Diagnosisarena: Benchmarking diagnostic reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 3074–3098.
- Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

A Case Study

This appendix provides the complete details of the case study discussed in Section 4.5.

B Prompt Templates

The prompt templates used for evaluating the reasoning quality of medical LLMs and for inference are shown in Tables 6 and 7, respectively.

C Implementation Details

In this study, we choose Llama3.1-8B-Instruct¹ and Qwen2.5-7B-Instruct² as the backbone of BELIEF. We choose DiagRL-Corpus³ as our retrieval Corpus. We initiated the training process with Supervised Fine-Tuning (SFT) as a cold start, followed by Reinforcement Learning (RL). Five reserved tokens (<search>, </search>, <results>, </results>, <wait>) are appended to the tokenizer with embeddings and LM-head rows initialized to the mean of existing vectors. All the experiments are conducted on an A100 NVIDIA GPU with 80G graphic memory in a Linux environment. The candidate set $\mathcal{C}(x)$ is formed by de-duplicating the final answers of $N = 8$ rollouts (inserting y^* if absent), yielding on average $K = 10$ candidates; trajectories are partitioned into $\Delta = 128$ -token chunks, and $p_k(t)$ is evaluated under the fixed cue $q_{\text{ans}} = \text{"Based on the above reasoning, the final answer is"}$. During training, the ground-truth diagnosis y^* is appended to the candidate set if it is absent, ensuring a valid target for BBUM. To avoid label leakage, y^* is never introduced at inference, where the model directly generates reasoning, retrieval actions, and final diagnoses without explicitly constructing the candidate set, posterior distribution, or entropy.

BBUM computes ρ_k as the cosine similarity between BGE⁴ embeddings of c_k and y^* , with $\theta_\rho = 0.7$. DERM uses entropy threshold $\eta = 0.6 \log k$, search-reward scale $\lambda_r = 1$, over-querying penalty $\beta = 0.1$, budget $M = 2$, and top- $k = 5$ passages retrieved by MedEmbed (Balachandran, 2024)⁵ with MMR diversity reranking (Carbonell and

α	0	0.2	0.4	0.6	0.8
Acc@1	57.4	58.1	58.0	58.3	57.6
Avg. searches	0.45	0.44	0.46	0.42	0.36

Table 4: Sensitivity of BELIEF-Llama3.1-8B to the training-time entropy threshold $\eta = \alpha \log K$ on MIMIC-C. Avg. searches denotes the average number of inference-time retrieval actions per case. The default $\alpha = 0.6$ is shown in bold.

Goldstein, 1998) ($\lambda_{\text{MMR}} = 0.5$) from a clinical corpus. The rethink reward uses $\delta = 0.2$, and R_{out} is returned by a regex format checker combined with BGE semantic similarity to y^* . Optimization uses clip $\epsilon = 0.2$, KL coefficient $\beta_{\text{KL}} = 0.01$, AdamW with peak learning rate 5×10^{-6} , generation temperature 0.9 with up to 2048 new tokens per trajectory.

D Entropy Threshold Sensitivity

We study the sensitivity of BELIEF to the training-time entropy threshold. Using BELIEF-Llama3.1-8B on MIMIC-C, we vary $\eta = \alpha \log K$, where $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8\}$, while keeping all other settings fixed. Table 4 reports diagnostic accuracy and average inference-time searches per case.

At the default $\alpha = 0.6$, BELIEF achieves the highest Acc@1 of 58.3 with 0.42 searches per case. Across the tested thresholds, Acc@1 remains within a narrow range of 57.4–58.3, demonstrating stable diagnostic performance across different threshold settings. Relative to the lower thresholds ($\alpha \leq 0.4$), $\alpha = 0.6$ improves Acc@1 while slightly reducing retrieval frequency. Increasing α to 0.8 further lowers retrieval frequency from 0.42 to 0.36 searches per case, but decreases Acc@1 by 0.7 points. These results identify $\alpha = 0.6$ as a favorable balance between diagnostic accuracy and retrieval cost.

E Cross-model hallucination & calibration comparison

Hallucination metrics. To quantify hallucination behaviour we report three complementary metrics (lower is better for all of them), corresponding to the three panels of Figure 2.

(1) Vocabulary hallucination rate (strict-and-wrong) measures how often a model invents a diagnosis that is both *outside the official disease vo-*

¹<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

²<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

³<https://huggingface.co/datasets/QiaoyuZheng/DiagRL-Corpus>

⁴<https://huggingface.co/BAAI/bge-large-en>

⁵<https://huggingface.co/abhinand/MedEmbed-large-v0.1>

cabulary $\mathcal{V}$ and incorrect:

$$\text{VHR}_{\text{strict_wrong}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i \notin \mathcal{V} \wedge \hat{y}_i \neq y_i^*],$$

where $\hat{y}_i$ is the predicted disease (after canonical normalisation), y_i^* the ground-truth label set, and N the number of evaluation samples. This metric isolates the *dangerous* hallucinations – confidently named diseases that are neither real nor correct.

(2) Expected calibration error (ECE) measures the gap between predicted confidence and empirical accuracy. With M equal-width confidence bins $\{B_m\}_{m=1}^M$,

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|,$$

where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ denote the average accuracy and mean predicted confidence within bin B_m . A well-calibrated model satisfies $\text{acc}(B_m) \approx \text{conf}(B_m)$, and is unlikely to confidently emit hallucinations.

(3) Area under the risk–coverage curve (AURC) captures selective-prediction quality. Sorting all samples by descending confidence and letting $\text{cov}(\tau)$ and $\text{risk}(\tau)$ denote the coverage and error rate at threshold τ ,

$$\text{AURC} = \int_0^1 \text{risk}(\text{cov}^{-1}(c)) dc.$$

A low AURC means the model’s own confidence reliably tells us *when* it is about to hallucinate, so a downstream system can defer the most risky cases.

Symptoms	A 57-year-old male presented with a widespread petechial rash that had been present for 3 days. His medical history included hypertension, which was managed with lisinopril. He reported experiencing malaise, fever, and chills 3 weeks prior to his arrival, but these symptoms had resolved by the time he presented. In the emergency department, a chest radiograph revealed minimal perihilar interstitial opacities without any focal lobar consolidation. The findings were consistent with conditions such as atypical pneumonia, pulmonary hemorrhage, or pulmonary edema. The patient did not exhibit hemoptysis, leukocytosis, or other signs indicative of acute heart failure. Additionally, he had no respiratory or neurological complaints and did not require oxygen supplementation during his hospital stay. Laboratory results showed a normal haptoglobin level and an unremarkable peripheral blood smear. Test results: Initial laboratory results showed a platelet count of $4,000\mu L$ with no other concerning complete blood count abnormalities. The patient was screened for SARS-CoV-2 infection and diagnosed with COVID-19 pneumonia.
Llama3.1-8B	The patient's presentation with a widespread petechial rash, which developed 3 weeks after his initial symptoms of malaise, fever, and chills, is consistent with TTP. The normal haptoglobin level and unremarkable peripheral blood smear rule out hemolytic anemia, which is often seen in TTP. The patient's platelet count of $4,000\mu L$ is low, but not critically low, which is also consistent with TTP. The presence of COVID-19 pneumonia does not exclude TTP, as both conditions can occur simultaneously. The patient's lack of respiratory or neurological complaints and normal chest radiograph findings further support the diagnosis of TTP. The patient's symptoms and laboratory findings strongly suggest Thrombotic Thrombocytopenic Purpura (TTP).
Ours	The patient presents with a petechial rash, thrombocytopenia ($4,000\mu L$), and a recent history of COVID-19 pneumonia. The key clinical features—subcutaneous hemorrhages, low platelets, and infectious etiology—narrow the differential diagnosis to viral-induced thrombocytopenic purpura. <search>petechial rash + thrombocytopenia + recent viral infection</search> [1. Patient Record: Immune Thrombocytopenic Purpura. phenotype: Hemoptysis, Rash, Petechiae, Oral blisters 2. Guideline: Scrub typhus and dengue virus infection (co-infection). Symptoms: Fever, headache, myalgia, rash, thrombocytopenia, lymphadenopathy, eschar, hepatomegaly, splenomegaly, nausea, vomiting, fatigue, conjunctival injection 3. Guideline: HIV associated thrombocytopenia. Symptoms: Thrombocytopenia, Bleeding episodes, Petechiae, Ecchymoses, Fatigue, Anemia, Neutropenia, Hypersplenism 4. Guideline: Hepatitis C virus infection with severe cutaneous vasculitis. Symptoms: Cutaneous vasculitis, Purpura, Skin ulcer, Arthralgia, Fatigue, Fever, Peripheral neuropathy, Raynaud phenomenon 5. Guideline: Parvovirus Infection. Symptoms: Fever, Headache, Cough, Sore throat, Slapped cheek rash, Lacy rash on trunk and limbs, Joint pain, Joint swelling, Itching, Diarrhea, Runny nose, Upset stomach, Anemia] <results> Result 2 (Scrub typhus/dengue co-infection) is unlikely due to lack of Scrub typhus risk factors or dengue-specific signs. Result 3 highlights HIV-associated thrombocytopenia and opportunistic infections, but the patient's HIV status is not mentioned. Result 4 describes parvovirus-induced vasculitis, which typically presents with a lacy rash and joint pain, not petechiae. Result 5 notes parvovirus can cause anemia, but the patient's normal haptoglobin and peripheral smear exclude hemolysis. The most plausible explanation is immune-mediated platelet destruction secondary to COVID-19, as viral infections are known to induce immune thrombocytopenia (ITP). The absence of fever, leukocytosis, or organ dysfunction further supports a self-limiting process rather than sepsis or acute lung injury. </results> The patient's thrombocytopenia and petechial rash are consistent with immune thrombocytopenic purpura (ITP), particularly in the context of a recent viral infection (SARS-CoV-2). The normal haptoglobin and peripheral smear rule out acute hemolysis, while the absence of systemic symptoms (e.g., fever, leukocytosis) reduces suspicion for bacterial pneumonia or sepsis. The differential diagnosis includes HIV-associated thrombocytopenia and parvovirus vasculitis, but the clinical context and lack of risk factors for HIV or dengue make ITP the most likely diagnosis. The patient's symptoms and laboratory findings strongly suggest Immune Thrombocytopenic Purpura (ITP).

Table 5: **A case study of clinical diagnosis.** The retrieved content is highlighted in yellow, and the final diagnostic results are enclosed in boxes.

evaluation prompt	Given a medical problem and a response from a large language model, please evaluate the response based on the following criteria: - Logical Coherence: The reasoning must be internally consistent, following a rational chain of thought from evidence to conclusion without contradictions. - Factual Accuracy: The medical knowledge, including pathophysiology and diagnostic criteria, must be scientifically correct and grounded in established clinical facts. - Evidence Faithfulness: The reasoning must stay strictly within the context of the provided case, avoiding any fabricated information or "hallucinations." - Interpretability & Clarity: The response must be transparent, well-organized, and easy for a clinical expert to follow, using precise medical terminology. - Information Utilization: The analysis should integrate all relevant data from the case, including key positive findings and pertinent negative findings. For each of the above five criteria, give a separate score from 0 to 10, where 0 means completely flawed and 10 means perfect. ## Output format ## Return ONLY a single JSON object, with no additional explanation, exactly in the following structure: "Logical Coherence": <number from 0 to 10>, "Factual Accuracy": <number from 0 to 10>, "Evidence Faithfulness": <number from 0 to 10>, "Interpretability & Clarity": <number from 0 to 10>, "Information Utilization": <number from 0 to 10> ##Here is the medical problem and the response from the large language model: Question: {{question}} Response: {{response}}
--------------------------	---

Table 6: Prompt for reasoning quality evaluation.

system prompt	You are a medical expert with advanced training in clinical reasoning and diagnostic medicine. Based solely on the patient's reported symptoms and relevant clinical information, determine the most likely diagnosis. Provide five possible disease diagnoses, each diagnosis must be presented only inside the box <code>\boxed{ }</code>. Do not include any explanations, qualifiers, or additional text inside the box. Example outputs: <code>\boxed{ {diagnosis 1} }, \boxed{ {diagnosis 2} }, \boxed{ {diagnosis 3} }, \boxed{ {diagnosis 4} }, \boxed{ {diagnosis 5} }</code>
----------------------	---

Table 7: System prompt for inference.