WHEN ANSWERS AGREE BUT REASONING DISAGREES: CONFLICT-AWARE AGENTIC RAG FOR MEDICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) show promise in medical question answering, yet extended reasoning can introduce factual errors and conflicting claims that undermine answer reliability. Recent multi-round retrieval-augmented generation (MA-RAG) mitigates these issues through dynamic retrieval using answer-level consistency signals, leading to state-of-the-art performance yet overlooking the conflicts within individual reasoning trajectories. Specifically, it will cost N times tokens to get the retrieval signal during answer sampling and fails when the model consistently produces incorrect answers through disordered reasoning. In this paper, we begin by analysing reasoning trajectories and explore how they can guide retrieval and repair direction. We observe that trajectories can be identified into three patterns: smooth, repeat and conflict. Notably, trajectories containing conflicting claims have substantially lower answer accuracy than repetitive or smooth reasoning, suggesting that these semantic patterns offer a useful signal for external refinement. Motivated by this finding, we propose a **C**onflict-aware **A**gentic RAG framework for medical **RE**asoning (**CARE**), which allows for a more reliable reasoning process with fewer tokens by analyzing reasoning trajectories to guide targeted interventions. It is achieved by preserving smooth reasoning, truncating repetition, and retrieving evidence to resolve factual conflicts, then integrating relevant reasoning and evidence to produce more reliable answers. Experiments on eight medical benchmarks show that CARE improves average accuracy by **2%** over the strong adaptive RAG baseline MA-RAG while reducing generated tokens by approximately **77%**. On MedThink-Bench, CARE further improves alignment with expert reasoning by **12.81 points** across four semantic metrics. These results demonstrate the potential of trajectory-level conflict detection for improving both the reliability and efficiency of medical reasoning. Our code is available at https://anonymous.4open.science/r/CARE-C3F1/.

1 INTRODUCTION

Complex medical question-answering tasks often require both broad medical knowledge and rigorous logical reasoning (Zhou et al., 2025; Jiang et al., 2026). Large language models (LLMs) have made major progress in natural language understanding and multi-step reasoning, achieving strong performance across a wide range of downstream tasks (Achiam et al., 2023). To further improve their ability to solve complex reasoning problems, post-training methods, such as reinforcement learning, have recently been developed to encourage models to produce longer and more complex reasoning processes (Shao et al., 2024; Yang et al., 2025; Guo et al., 2025). These advances have increased the potential of LLMs in medical applications. However, LLMs' long reasoning does not always lead to more reliable answers. A reasoning trajectory may contain incorrect medical facts or conflicting claims, which can undermine the final answer (Chen et al., 2025b; Pandit et al., 2025).

Retrieval-augmented generation (RAG) improves reliability by introducing external medical evidence (Lewis et al., 2020; Xiong et al., 2024). Traditional RAG retrieves documents only once based on the original question, which is often insufficient for complex reasoning and may introduce noise (Jiang et al., 2023; Asai et al., 2024). To tackle this, representative adaptive retrieval mechanisms support multi-round retrieval by dynamically determining when and what to trigger external

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

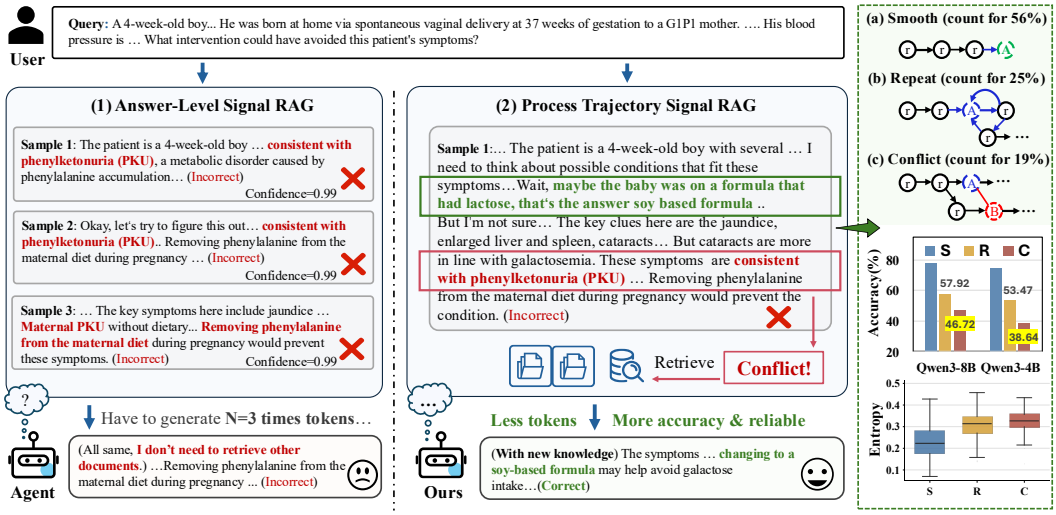


Figure 1: Our observation: Answer-level consistency can mask reasoning errors. **Left**: multiple sampled responses converge on the same incorrect answer, leaving an answer-level retrieval signal unable to detect the need for additional evidence. **Middle**: conflicting claims within a single trajectory guide targeted retrieval and reasoning repair. **Right**: three trajectory patterns with their associated answer accuracy and sequence entropy. The accuracy and entropy metrics are calculated on eight datasets, and the entropy presents the results for the Qwen3-4B model.

retrieval, achieving state-of-the-art (SOTA) results (Jeong et al., 2024; Jiang et al., 2025; Wu et al., 2026). But they usually rely on token-level confidence (Jiang et al., 2023) or answer-level consistency (Jiang et al., 2025; Wu et al., 2026). Confidence is unreliable because LLMs may produce high-confidence hallucinations (Savage et al., 2025), while answer-level signals overlook conflicts *within* an individual reasoning trajectory. As illustrated in Figure 1, multiple reasoning paths may reach the same incorrect answer, and the answer-level consistency signal therefore misses the need for additional evidence, even when the reasoning itself contains conflicting claims.

This observation motivates us to examine common patterns in individual reasoning trajectories and explore how they can guide retrieval and reasoning repair. As shown in Figure 1, we identify three reasoning patterns: *smooth*, *repeat*, and *conflict*, by diagnosing their trajectory graph. We find that these three patterns have answer accuracies of approximately 75%, 55%, and 42%, respectively. We also measure sentence entropy for each pattern, where higher values indicate greater uncertainty during reasoning generation (Farquhar et al., 2024). Notably, repetitive and conflicting trajectories show similar entropy levels despite their substantial difference in answer accuracy. This contrast suggests that entropy alone does not distinguish these answers well, whereas trajectory semantic patterns provide a more direct signal of claims requiring external verification. However, it remains challenging to construct effective queries and properly use the retrieved documents. In addition, given the large number of reasoning steps and evidence documents, these sources must be effectively consolidated to produce a reliable answer.

To address these challenges, we propose a **C**onflict-aware **A**gentic RAG framework for medical **RE**asoning (**CARE**), which achieves a more reliable reasoning process with fewer tokens by analyzing reasoning trajectories to guide targeted interventions. Concretely, our framework consists of three agents: i) A Solver Agent generates an initial retrieval-augmented response. ii) A Critic Agent then diagnoses its reasoning pattern by constructing a trajectory graph and applies a corresponding intervention. It forwards smooth reasoning responses, truncates repetitive reasoning, or conducts targeted retrieval for conflicting judgments. In the conflict branch, a Selector jointly evaluates retrieved documents and the model-generated response to select question-relevant context. iii) Finally, the Repair Agent combines the processed response and available evidence into the final answer. Experiments on eight medical benchmarks show that CARE improves average accuracy by 2% over the strong adaptive RAG method (MA-RAG) while reducing generated tokens by approximately 77%. On MedThink-Bench, CARE also improves the semantic and textual consistency between generated and expert reasoning by 12.81 points across four semantic and textual metrics.

In sum, our main contributions are as follows:

- To the best of our knowledge, we introduce the first trajectory-level retrieval signal based on factual semantic patterns within a single reasoning process and provide an effective and efficient approach to improving the reliability of LLM-based medical reasoning.
- We propose CARE, an agentic framework for medical reasoning based on trajectory graph analysis that identifies and handles several reasoning issues through targeted interventions.
- We demonstrate consistent gains across eight benchmarks, improving accuracy by an average of 2% over strong RAG baselines and alignment with expert-annotated reasoning by 12.81 points, with efficiency and scalability.

2 Related Work

Long-chain reasoning in large language models. Chain-of-thought (CoT) prompting improves LLM reasoning by explicitly generating intermediate steps (Wei et al., 2022; Kojima et al., 2022). Recent reasoning-oriented models further extend this paradigm with long-form inference. For example, DeepSeek-R1 uses reinforcement learning to induce self-reflection and verification (Guo et al., 2025), while Qwen3 supports both *thinking* and *non-thinking* modes for different reasoning demands (Yang et al., 2025). Although longer reasoning often improves performance on complex tasks, it can also introduce redundant reasoning and error accumulation (Chen et al., 2025b; He et al., 2025). In particular, the reasoning trajectory may contain repetitive or conflicting claims even when the final answer appears plausible. Our work therefore focuses on trajectory-level reasoning patterns and uses such patterns as signals for targeted intervention.

Retrieval-augmented generation for medical reasoning. RAG is widely used in medical question answering to ground LLM outputs in external knowledge (Lewis et al., 2020; Xiong et al., 2024). Conventional RAG typically retrieves evidence once from the original question, which may be insufficient for complex multi-step reasoning (Jiang et al., 2023; 2025). Recent methods therefore adopt iterative retrieval. Adaptive-RAG triggers retrieval via query complexity classification Jeong et al. (2024), TC-RAG follows with state-managed processing Jiang et al. (2025), while MA-RAG uses disagreement among multiple candidate responses to guide multi-round retrieval (Wu et al., 2026). However, such inter-response signals may fail when generating consistency-error answers and require generating multiple long responses. CARE instead detects *intra-trajectory patterns* and selectively performs truncation or targeted retrieval only when needed.

3 Method

Key Definitions. Given a medical query q and a medical corpus $\mathcal{C}$, CARE produces the final answer $\hat{y}$ through the collaboration of three agents. Figure 2 provides an overview of the CARE framework. We represent the inference state of CARE as

$$\mathcal{S}_q = \{D_q, T_q, G_q, z_q, E_q\}, \tag{1}$$

where D_q denotes the initially retrieved evidence, T_q the long-form reasoning trajectory, and G_q the corresponding heterogeneous trajectory graph. The pattern variable $P \in \{P_s, P_r, P_c\}$ denotes the `smooth`, `repeat`, and `conflict` pattern, respectively. For pattern P_c, E_q denotes the additional evidence selected to resolve the competing answer commitments.

3.1 Solver Agent: Evidence-Grounded Reasoning

The Solver Agent serves as the primary reasoning engine of CARE. It first retrieves medical evidence relevant to q and then generates a reasoning trajectory conditioned on the retrieved context. Given the Solver prompt $\mathbf{P}_{\text{solver}}$, the query, and the retrieved documents from the medical corpus ($\mathcal{C}$): $D_q = \text{Retrieval}(q, \mathcal{C})$, the Solver Agent generates

$$T_q = \mathbf{LLM}\left(\mathbf{P}_{\text{solver}}, q, D_q\right). \tag{2}$$

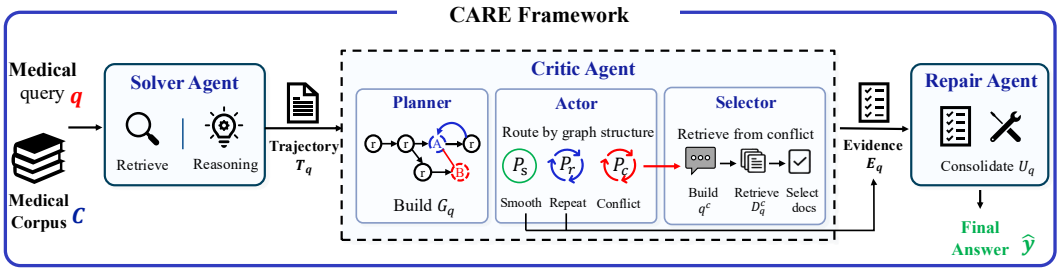


Figure 2: Overview of CARE. **The Solver** generates a reasoning trajectory T_q. Within the **Critic**, the Planner constructs G_q, and the Actor diagnoses the Smooth, Repeat, or Conflict pattern (P_s, P_r, or P_c), while P_c activates the Selector to retrieve evidence and jointly select effective evidence from retrieved documents and T_q. **Repair Agent** consolidates all evidence U_q to produce answer $\hat{y}$.

Once we obtain the reasoning trajectories, a key challenge is identifying distinct patterns within them. We propose constructing heterogeneous trajectory graphs to facilitate this process. Specifically, CARE treats each sentence as a node of the trajectory graph, and we denote the resulting reasoning sequence as $T_q = (r_1, \ldots, r_n)$, which is subsequently organized by the Critic Agent.

3.2 Critic Agent: Graph-Based Diagnosis and Evidence Acquisition

The Critic Agent transforms the Solver trajectory into an actionable reasoning state through three internal roles: **Planner**, **Actor**, and **Selector**. The Planner constructs the trajectory graph, then Actor identifies its structure and determines the pattern, a Selector acquires targeted evidence for the conflict pattern P_c, which needs to use other evidence to clarify the original answer.

3.2.1 Planner: Heterogeneous Trajectory Graph

The Planner maps a reasoning trajectory T_q to an undirected heterogeneous graph G_q with two node types: reasoning nodes V_R for intermediate reasoning states and answer nodes V_A for answer commitments. It contains three edge types: trajectory edges E_{traj} that preserve reasoning progression, support edges E_{support} that link reasoning nodes to answer nodes, and conflict edges E_{conflict} that connect incompatible answers. Formally,

$$G_q = f_G(T_q) = (V_R \cup V_A, E_{\text{traj}} \cup E_{\text{support}} \cup E_{\text{conflict}}) . \tag{3}$$

Repeated commitments to the same answer share a canonical answer node, forming a *repeat cycle* when a later reasoning path reconnects to it. A *conflict cycle* arises when paths from a shared upstream state reach distinct answer nodes connected by E_{conflict}. These cycles capture repeated and conflicting answer commitments, respectively. The rule-based function f_G constructs G_q in a single pass over T_q, with $O(N)$ time complexity for $N = |T_q|$, introducing negligible time overhead.

3.2.2 Actor: Graph-Based Pattern Diagnosis

The Actor determines the inference pattern from the graph structure of G_q:

$$z_q = f_A(G_q) \in P_s, P_r, P_c. \tag{4}$$

The first graph cycle to close determines the pattern decision: a conflict cycle induces P_c, while a repeat cycle corresponds to P_r. If both cycles close at the same position, P_c takes precedence. If no cycle is formed before the trajectory terminates, the Actor assigns P_s. For P_r, CARE retains the trajectory prefix $T_q^{(k)}$ ending at the k-th repeat-cycle closure. This design is motivated by self-consistency (Wang et al., 2023): when multiple reasoning paths converge to the same answer, their agreement provides stronger evidence for that prediction. For P_c, CARE instead invokes conflict-guided evidence acquisition.

3.2.3 Selector: Conflict-Aware Evidence Selection

The conflict pattern identifies the competing answer candidates that define the retrieval target. Let c^- and c^+ denote the two answers associated with the detected conflict cycle. We use them to

construct the retrieval query:

$$\text{query}^{\text{c}} = \phi\left(q, c^{-}, c^{+}\right), \tag{5}$$

where ϕ is the template that expresses the medical distinction between the competing candidates. Then we get the candidate evidence set $D_q^{\text{c}} = \text{Retrieval}(\text{query}^{\text{c}}, \mathcal{C})$. Because the model has uncertain knowledge about conflicting content, as we mentioned in Section 1, it is more vulnerable to noise in the retrieved documents. To address this issue, we introduce a Selector to filter the retrieved evidence. In particular, we treat the model's original output T_q as an additional knowledge document and jointly consider it with the retrieved documents during evidence selection:

$$(a_q, E_q) = \mathbf{LLM}\left(\mathbf{P}_{\text{selector}}, q, \{D_q^{\text{c}}, T_q\}\right), \tag{6}$$

where a_q denotes the analysis process and E_q the selected documents. The resulting (a_q, E_q) and other pattern outputs will be forwarded to the Repair Agent.

3.3 Repair Agent: Pattern-Conditioned Evidence Consolidation

As discussed in Section 1, long reasoning trajectories may contain hesitation or conflicting intermediate conclusions, reducing the reliability of the final reasoning. We therefore introduce a Repair Agent to consolidate the preceding reasoning and evidence into a coherent final prediction.

Specifically, we denote all the context from the Critic Agent by U_q, where the smooth pattern P_s uses the full trajectory T_q, P_r uses the truncated trajectory T_q^{k}, and P_c uses the Selector output (a_q, E_q). Given the Repair prompt $\mathbf{P}_{\text{repair}}$, the final prediction is

$$\hat{y} = \mathbf{LLM}\left(\mathbf{P}_{\text{repair}}, q, U_q\right). \tag{7}$$

3.4 Theoretical Analysis

We analyze CARE as a paired intervention on the same initial Solver trajectories. Let Y_0 and Y_1 denote the predictions before and after the intervention under study, respectively, and use Y^{+} and Y^{-} to denote correct and incorrect predictions. For trajectory type $z \in \{\text{f}, \text{r}, \text{c}\}$, define

$$p_z = \Pr(Y_0^{-} \mid z), \qquad g_z = \Pr(Y_1^{+} \mid Y_0^{-}, z), \qquad h_z = \Pr(Y_1^{-} \mid Y_0^{+}, z),$$

where p_z is the pre-intervention error rate, g_z the correction rate, and h_z the regression rate.

Proposition. The accuracy gain for trajectory type z is

$$\Delta_z = p_z g_z - (1 - p_z) h_z. \tag{8}$$

Hence, the intervention improves expected accuracy whenever $p_z g_z > (1 - p_z) h_z$, where $p_z g_z$ represents the correction gain and $(1 - p_z) h_z$ the regression loss.

Aggregating over trajectory types gives

$$\Delta_{\text{CARE}} = \sum_z \Pr(z) \left[p_z g_z - (1 - p_z) h_z\right]. \tag{9}$$

For the full CARE framework, Y_0 is the initial Solver prediction and Y_1 is the final CARE prediction. To isolate the contribution of conflict-guided retrieval, Y_0 is obtained with the same Repair Agent but without conflict-specific evidence, while Y_1 additionally uses E_q.

4 Experiments

4.1 Setup

Datasets. We evaluate CARE on eight medical question-answering benchmarks covering diverse medical knowledge and reasoning requirements: MMLU-P (MMLU-Pro) (Wang et al., 2024), MedB (MedBullets) (Chen et al., 2025a), MedE (MedExpQA) Alonso et al. (2024), MedMC (MedMCQA) (Pal et al., 2022), MedQA (Jin et al., 2021), MedX (MedXpertQA) (Zuo et al., 2025), MMLU-M (MMLU-Medical) (Hendrycks et al., 2021), and MedThink (Medthink-Bench) (Zhou et al., 2025).

Table 1: Accuracy (%) across eight medical question-answering benchmarks. **Avg.** denotes the unweighted average. **Bold** and underlined indicate the best and second-best results for each backbone.

Method	MMLU-P	MedB	MedE	MedMC	MedQA	MedX	MMLU-M	MedThink	Avg.
Backbone LLM: Qwen3-4B									
Base	49.51	40.58	63.20	54.17	58.68	13.27	74.38	30.60	48.05
CoT	51.59	43.83	62.40	55.39	61.43	13.06	74.93	29.60	49.03
CoT-Think	64.06	**50.65**	68.80	57.78	69.05	14.49	82.64	34.40	55.23
Single-RAG	61.00	47.08	**74.40**	63.45	69.29	18.37	82.83	36.60	56.63
Adaptive-RAG	61.61	44.81	72.80	63.59	70.15	18.37	82.00	36.40	56.22
TC-RAG	61.12	46.43	71.20	64.86	68.42	18.98	82.74	38.00	56.47
MA-RAG	62.47	49.68	67.20	62.28	**72.19**	15.92	83.38	34.40	55.94
CARE	65.65	48.70	72.00	66.36	70.31	**18.98**	84.76	39.20	58.25
CARE+2Rounds	**66.38**	50.00	72.00	**66.94**	71.33	18.16	**84.94**	**39.40**	**58.64**
Backbone LLM: Qwen3-8B									
Base	58.68	47.73	64.80	58.79	64.18	13.27	79.98	32.60	52.50
CoT	61.98	49.35	66.40	57.78	67.24	14.29	79.98	34.60	53.95
CoT-Think	69.56	**57.79**	73.60	64.55	77.22	16.33	86.41	40.40	60.73
Single-RAG	67.36	57.14	74.40	66.75	77.45	20.82	86.78	41.60	61.54
Adaptive-RAG	66.63	54.87	72.00	66.91	76.59	20.41	86.32	41.20	60.62
TC-RAG	67.85	53.90	73.60	66.75	76.98	20.61	86.41	41.40	60.94
MA-RAG	**70.78**	57.14	76.00	65.41	78.32	17.96	86.41	42.00	61.75
CARE	69.32	56.82	74.40	**68.78**	77.77	21.84	87.88	**44.40**	62.65
CARE+2Rounds	70.29	56.17	**76.00**	68.75	**78.40**	**22.45**	**87.97**	44.00	**63.00**

We report accuracy on all benchmarks, and we evaluate reasoning reliability on MedThink using METEOR (Denkowski & Lavie, 2014), BERTScore F1 (Zhang et al., 2020), BLEU (Chen & Cherry, 2014), and ROUGE-L F1 (Lin, 2004; Guo et al., 2024) metrics to measure the semantic alignment between model-generated reasoning and human expert-annotated reasoning. Details about dataset statistics and metric descriptions can be found in Appendix C.1.

Baselines. We compare CARE against three non-retrieval baselines, including (1) Base, (2) CoT (Wei et al., 2022), and (3) CoT-Think (Yang et al., 2025), as well as four retrieval-augmented baselines: (1) Single-RAG (Lewis et al., 2020), (2) Adaptive-RAG (Jeong et al., 2024), (3) TC-RAG (Jiang et al., 2025), and (4) MA-RAG (Wu et al., 2026). These baselines cover both one-shot retrieval and adaptive or multi-round retrieval strategies. We conduct experiments with Qwen3-4B and Qwen3-8B (Yang et al., 2025) as the main backbone models, and further evaluate Qwen3.5-27B (Team, 2026) to study whether the proposed method remains effective as model capacity increases. Details about baseline descriptions and experimental settings are presented in Appendix D and E.

4.2 Results

Main Results. Table 1 shows that CARE consistently improves both backbones, where the same LLM model is used in all agents. *CARE+2Rounds* denotes that, for the conflict pattern, the output produced after Critic Agent processing is fed back into the Critic Agent as a new input, and the entire procedure is repeated once more. Our CARE can achieve an average accuracy improvement of $\tilde{2}\%$ over MA-RAG. Specifically, on Qwen3-4B, CARE and *CARE+2Rounds* achieve average accuracies of 58.25% and 58.64%, respectively, compared with 56.63% for the strongest competing RAG baseline. On Qwen3-8B, they reach 62.65% and 63.00%, outperforming MA-RAG at 61.75%. Gains are particularly clear on MedThink, where CARE improves CoT-Think from 34.40% to 39.20% for 4B and from 40.40% to 44.40% for 8B. Overall, trajectory-aware intervention provides consistent gains, while a second round yields smaller additional improvements.

Reasoning Reliability Beyond Answer Accuracy. CARE improves not only answer accuracy but also alignment with human-expert reasoning. On MedThink, it achieves the best accuracy among all

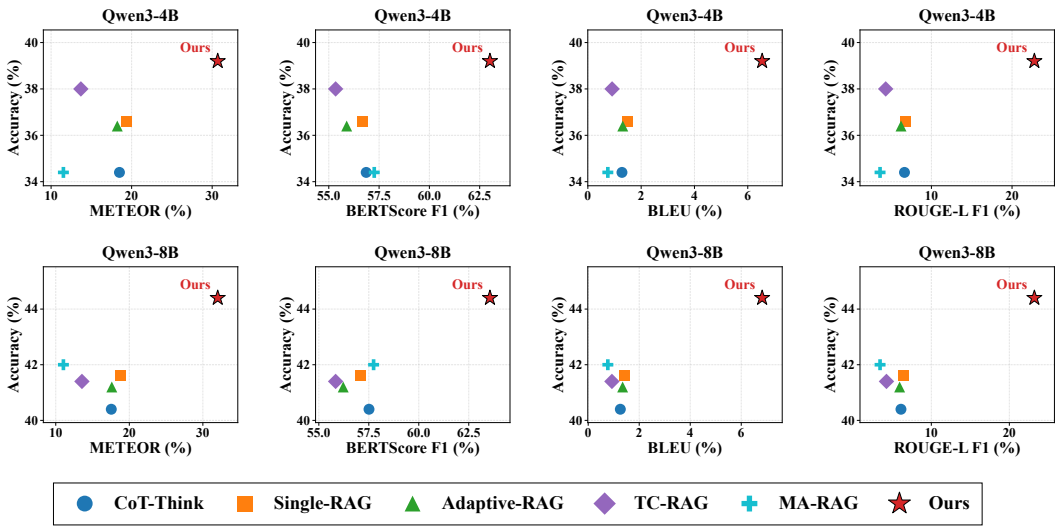


Figure 3: Accuracy versus reasoning-reliability metrics on Qwen3-4B and Qwen3-8B. CARE improves both answer accuracy and agreement with expert reasoning.

Table 2: Ablation study for reasoning reliability metrics on MedThink dataset.

Model	Stage	Accuracy	METEOR	BERTScore F1	BLEU	ROUGE-L F1
Qwen3-4B	CoT-Think	34.40%	18.49%	56.86%	1.28%	6.66%
	w. Initial Retrieval	37.40%	17.81%	56.65%	1.29%	6.48%
	w. Critic Agent	**39.20%**	23.07%	58.18%	2.66%	11.30%
	w. Repair Agent	**39.20%**	**30.69%**	**63.01%**	**6.54%**	**22.77%**
Qwen3-8B	CoT-Think	40.40%	17.55%	57.51%	1.26%	6.11%
	w. Initial Retrieval	42.00%	17.48%	56.89%	1.27%	6.18%
	w. Critic Agent	43.40%	22.72%	58.61%	2.63%	10.84%
	w. Repair Agent	**44.40%**	**32.02%**	**63.56%**	**6.82%**	**23.21%**

baselines and an average gain of **12.81 points** across four reasoning-alignment metrics and two backbone models (Figure 3); detailed results are reported in Appendix F.1. Table 2 shows that retrieval mainly improves accuracy, whereas the Critic and Repair stages substantially improve reasoning alignment. For Qwen3-4B, the Critic and Repair stages achieve similar accuracy, yet Repair increases METEOR from 23.07% to 30.69% and BERTScore F1 from 58.18% to 63.01%. Qwen3-8B shows the same trend, indicating that answer accuracy alone cannot fully reflect reasoning quality.

Validation of the Gain Condition. We evaluate the sufficient condition from Section 3.4,

$$p_z g_z > (1 - p_z) h_z,$$

where p_z is the initial error probability, g_z the correction probability for incorrect predictions, and h_z the corruption probability for correct predictions. Table 3 shows positive gains for all trajectory patterns. Conflict trajectories achieve the largest gain (+4.93 points), with the highest initial error ($p_z = 61.36\%$) and correction probability ($g_z = 14.02\%$), leading to the largest observed gain. These results are consistent with our theoretical analysis and provide empirical support for using trajectory-pattern based interventions in CARE.

Ablation Study. We conduct an incremental ablation study on Qwen3-4B to quantify the contribution of each CARE component, with Qwen3-8B results reported in the Appendix. Starting from the base reasoning model, we progressively add initial retrieval, the Critic Agent, and the Repair Agent. As shown in Figure 4, initial retrieval can degrade performance on some datasets, possibly because retrieved evidence conflicts with the model's internal knowledge and introduces reasoning instability. Performance improves consistently as the Critic and Repair components are added, highlighting the benefit of conflict-aware intervention and evidence consolidation.

Table 3: Empirical validation of the theoretical gain condition on Qwen3-4B over eight benchmarks.

Trajectory	#Samples	Y_0 Acc.	Y_1 Acc.	p_z	g_z	h_z	Gain
Complete	4,836	74.32	75.50	25.68	4.67	0.03	**+1.18**
Repeat	2,265	53.47	56.20	46.53	9.20	2.89	**+2.74**
Conflict	1,685	38.64	43.56	61.36	14.02	9.52	**+4.93**
Overall	8,786	62.10	64.40	37.90	9.01	1.80	**+2.30**

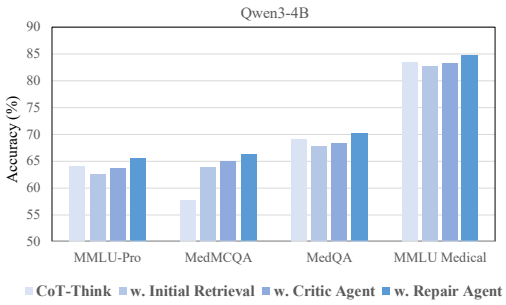


Figure 4: Ablation study of CARE's agents.

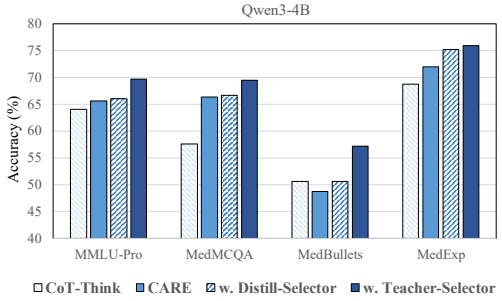


Figure 5: Analysis of Selector in Critic Agent.

In the standard CARE setting, the Selector uses the same backbone model as the other agents, which may limit its ability to judge its own reasoning and retrieved evidence. We therefore further study the effect of evidence selection by replacing the Selector with a stronger Qwen3.5-27B teacher while keeping all other components unchanged. As shown in Figure 5, a stronger Selector further improves reasoning performance. We additionally distill the teacher Selector using its selection trajectories on the MedQA (Jin et al., 2021) training set. The resulting distilled Selector also improves upon the standard CARE configuration, demonstrating the importance of high-quality evidence selection.

Trajectory Patterns versus Entropy. Figure 6 compares CARE's semantic trajectory patterns with sentence-level entropy, which is calculated as the average next-token entropy over all tokens in the reasoning content. Smooth trajectories generally have lower entropy, whereas Repeat and Conflict trajectories are more uncertain but substantially overlap. Entropy can therefore indicate difficulty without identifying its cause, while CARE distinguishes failure modes that require different interventions by semantic pattern analysis and can achieve effective improvements.

Inference Efficiency. We construct inference efficiency experiments, as presented in Table 4. The results show that CARE achieves higher accuracy with substantially less generation than multi-round RAG. Compared with MA-RAG, CARE reduces generated tokens by about 77% on both Qwen3-4B and Qwen3-8B. TC-RAG retrieves about 20 documents per question but remains less accurate than CARE. CARE therefore gains efficiency by selectively intervening on problematic trajectories rather than uniformly increasing retrieval or generation.

Table 4: Inference efficiency over eight datasets.

Model	Method	Docs	Gen.Tokens	Acc (%)
4B	Single-RAG	5.00	1,365.50	56.63
	Adaptive-RAG	4.99	1,555.90	56.22
	TC-RAG	19.67	2,389.40	56.47
	MA-RAG	6.57	6,500.84	55.94
	CARE	5.87	1,465.30	**58.25**
8B	Single-RAG	5.00	1,414.30	61.54
	Adaptive-RAG	4.99	1,613.90	60.62
	TC-RAG	19.98	2,457.20	60.94
	MA-RAG	7.07	7,180.67	61.75
	CARE	5.92	1,653.60	**62.65**

Repetitive Frequency Analysis. Figure 7 analyzes different cycle counts k as truncation signals for the *repeat* branch. Truncating the reasoning trajectory before $k = 3$ yields clear gains, while the improvement gradually saturates as k increases. We therefore set $k = 3$ in our experiments. Additional results are provided in the Appendix F.4.

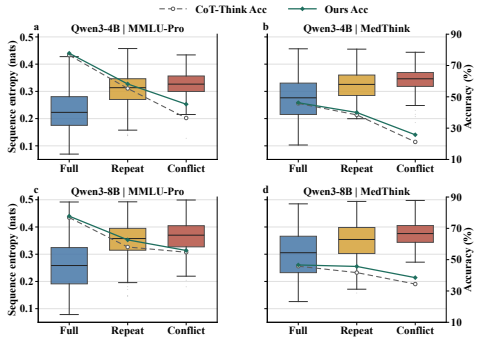


Figure 6: Entropy across reasoning patterns.

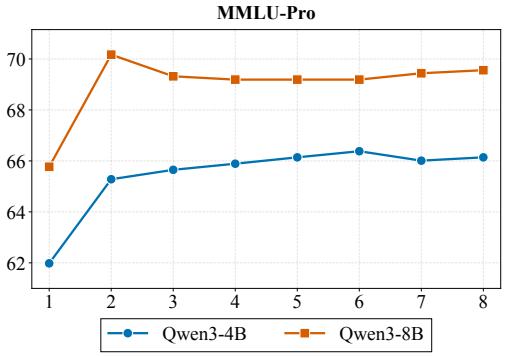


Figure 7: Repetition frequency analysis.

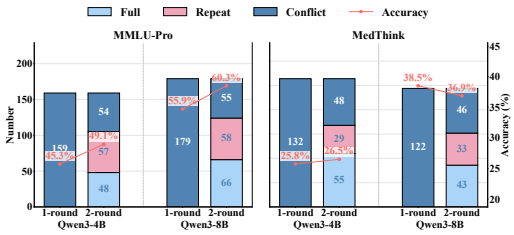


Figure 8: Conflict pattern transitions.

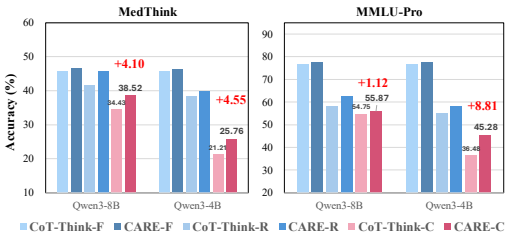


Figure 9: Improvements across patterns.

Can CARE Resolve Detected Conflicts? Figure 8 tracks the number of questions classified as the *conflict* pattern, where *1-round* denotes conflicts detected in the first CARE round and *2-round* those remaining in the second round. Our method substantially reduces conflicts in model outputs, improving both answer accuracy and reliability. Figure 9 further shows consistent gains across all patterns, with the largest improvement on the *conflict* branch. Additional results and analyses are provided in the Appendix F.3.

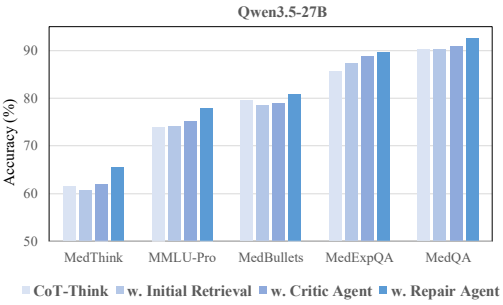


Figure 10: Ablation results on Qwen3.5-27B.

Scaling to a Larger Backbone. We also evaluate CARE on Qwen3.5-27B. As shown in Figure 10, the full pipeline improves over CoT-Think on all five benchmarks, with gains up to **+4.03** percentage points and an average improvement of about **+3.15 points**. Initial retrieval can occasionally hurt performance, including on MedBullets and MedThink, but the Critic and Repair stages recover these losses and produce final improvements. This suggests that trajectory-aware diagnosis remains useful even with a substantially stronger backbone.

5 Conclusion

We presented CARE, a conflict-aware agentic RAG framework that uses trajectory-level semantic patterns to guide targeted intervention in medical reasoning. By distinguishing smooth, repetitive, and conflicting trajectories, CARE selectively forwards, truncates, or retrieves additional evidence, followed by evidence consolidation through a Repair Agent. Experiments across eight medical benchmarks show that CARE improves both answer accuracy and reasoning reliability while substantially reducing generation cost. These results highlight trajectory-level diagnosis as an effective signal for building more reliable and efficient retrieval-augmented reasoning systems.

AI USE STATEMENT

We used AI tools to assist with language polishing during manuscript preparation and to help implement limited portions of the framework code. All research ideas, methodological designs, empirical findings, experimental analyses, and the main content of the paper were independently developed and validated by the authors.

REFERENCES

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

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 362–367, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3346. URL https://aclanthology.org/W14-3346/.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599, 2025a.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? on the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499. PMLR, 2025b. URL https://proceedings.mlr.press/v267/chen25bx.html.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL https://aclanthology.org/W14-3348/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

YiQiu Guo, Yuchen Yang, Ya Zhang, Yu Wang, and Yanfeng Wang. DictLLM: Harnessing key-value data structures with large language models for enhanced medical diagnostics. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10231–10241, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.609. URL https://aclanthology.org/2024.findings-acl.609/.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zy Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, et al. Can large language models detect errors in long chain-of-thought reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18468–18489, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 7036–7050, 2024.

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jiaming Shen, Jimeng Sun, and Jiawei Han. RAS: Retrieval-and-structuring for knowledge-intensive LLM generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fqqmeg61yd.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, et al. Tc–rag: Turing–complete rag's case study on medical llm systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11400–11426, 2025.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

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

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022.

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

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2356–2362, 2021.

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

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *arXiv preprint arXiv:2502.14302*, 2025.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H. Chen. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149, 2025. doi: 10.1093/jamia/ocae254. URL https://doi.org/10.1093/jamia/ocae254.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Wenhao Wu, Zhentao Tang, Yafu Li, Shixiong Kai, Mingxuan Yuan, Zhenhong Sun, Chunlin Chen, and Zhi Wang. From conflict to consensus: Boosting medical reasoning via multi-round agentic RAG. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=S7lpdz7NAW.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. Automating expert-level medical reasoning evaluation of large language models. *NPJ Digital Medicine*, 9(1):34, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80961–80990. PMLR, 13–19 Jul 2025.

APPENDIX CONTENTS

A Limitations and Future Work **14**

B Algorithm Pseudocode **14**

C Datasets and Evaluation Metrics **14**

C.1 Datasets 14

C.2 Evaluation Metrics. 15

D Baseline Methods **16**

D.1 Inference Methods without Retrieval. 16

D.2 Retrieval-Augmented Methods. 16

E Experimental Setup **17**

E.1 Model and Generation Settings 17

E.2 Retrieval Design and Knowledge Sources 17

F Supplementary Results and Analysis **17**

F.1 Detailed Reasoning Reliability Results 17

F.2 Ablation study 18

F.3 Pattern Analysis on the Full Dataset 19

F.4 Analysis of Different Repeat Times k 19

F.5 Retrieval for Repeat Patterns 20

G Case Study **21**

H Prompt Details **23**

A Limitations and Future Work

This work focuses on improving retrieval-augmented reasoning by analyzing reasoning trajectories to determine *when* retrieval is needed and *what* information should be retrieved. Despite the effectiveness of CARE, several limitations remain.

First, CARE still depends on the quality of retrieved evidence. Retrieved documents may be irrelevant, noisy, or inconsistent with the model's internal knowledge, making reliable evidence selection challenging. We mitigate this issue with the Selector in the Critic Agent, and our experiments show that a stronger Selector leads to better downstream performance. Designing more lightweight yet accurate evidence selectors is therefore an important direction for future work.

Second, we find that additional retrieval provides little benefit for questions exhibiting the *repeat* pattern (Appendix F.5). One possible explanation is that repeated convergence makes the model strongly committed to its current answer and less receptive to external evidence; another is that retrieved documents may introduce additional noise. Future work could explore better ways to organize and present external evidence, or train models to more effectively reconcile internal knowledge with retrieved information.

Finally, stronger models exhibit fewer *repeat* patterns, while *conflict* patterns remain present. This suggests that the distribution of reasoning failures may shift with model capability, but trajectory-level conflict detection remains relevant. Extending CARE to stronger models and broader domains, and studying how reasoning patterns evolve with model scale, are promising directions for future work. More broadly, reasoning trajectory patterns could also support self-evolving agents. For example, *repeat* and *conflict* patterns can serve as explicit signals for introducing external knowledge to correct uncertain or inconsistent reasoning. The corrected trajectories could then be incorporated into model training or persistent memory, enabling the agent to recursively refine its knowledge and reasoning behavior over time. Developing such trajectory-driven mechanisms for recursive self-evolution is an interesting direction for future research.

B Algorithm Pseudocode

Algorithm 1 summarizes the inference procedure of CARE. Given a medical query q, the Solver first retrieves relevant evidence from the medical corpus and generates a long-form reasoning trajectory T_q. The Planner then converts T_q into a heterogeneous trajectory graph G_q, based on which the Actor identifies the inference pattern according to the first cycle closure. For a smooth trajectory (P_s), CARE retains the full trajectory as the context for final consolidation. For a repeat trajectory (P_r), it retains only the prefix $T_q^{(k)}$ ending at the first repeat-cycle closure. For a conflict trajectory (P_c), the two competing answer commitments c^- and c^+ are used to formulate a conflict-specific retrieval query. The Selector jointly considers the newly retrieved documents and the original trajectory to identify effective evidence. Finally, the Repair Agent performs pattern-conditioned evidence consolidation over U_q and produces the final prediction $\hat{y}$.

C Datasets and Evaluation Metrics

C.1 Datasets

We evaluate CARE on 8,786 questions from eight medical multiple-choice benchmarks spanning foundational medical knowledge, clinical decision-making, expert-level reasoning, and reasoning-reliability evaluation with reference rationales. Detailed dataset statistics are summarized in Table 5.

MMLU-Pro (Wang et al., 2024) extends the original Massive Multitask Language Understanding (MMLU) benchmark with higher-quality questions and more competitive answer choices, providing a more rigorous evaluation of model capabilities. Our experiments utilize its health-domain subset, comprising 818 questions, to assess complex medical reasoning.

MedBullets (Chen et al., 2025a) is built from clinical practice questions published on the MedBullets educational website. We use its complete set of 308 five-option questions, which assess clinical diagnosis and management while requiring models to distinguish among multiple plausible distractors.

Algorithm 1 Inference Procedure of CARE

Require: Medical query q, medical corpus $\mathcal{C}$; Solver prompt $\mathbf{P}_{\text{solver}}$, Selector prompt $\mathbf{P}_{\text{selector}}$, Repair prompt $\mathbf{P}_{\text{repair}}$
Ensure: Final prediction $\hat{y}$

// Solver: evidence-grounded reasoning
$D_q \leftarrow \text{Retrieval}(q, \mathcal{C})$
$T_q \leftarrow \mathbf{LLM}(\mathbf{P}_{\text{solver}}, q, D_q)$
// Planner: trajectory graph construction
$G_q \leftarrow f_G(T_q)$ {Construct reasoning, answer, support, trajectory, and conflict relations}
// Actor: graph-based pattern diagnosis
$(z_q, k, c^-, c^+) \leftarrow f_A(G_q)$ {Diagnose by the first cycle closure}
if $z_q = P_s$ **then**
  $U_q \leftarrow T_q$
else if $z_q = P_r$ **then**
  $T_q^{(k)} \leftarrow \text{Prefix}(T_q, k)$
  $U_q \leftarrow T_q^{(k)}$
else if $z_q = P_c$ **then**
  // Selector: conflict-guided evidence acquisition
  $\text{query}^c \leftarrow \phi(q, c^-, c^+)$
  $D_q^c \leftarrow \text{Retrieval}(\text{query}^c, \mathcal{C})$
  $(a_q, E_q) \leftarrow \mathbf{LLM}\big(\mathbf{P}_{\text{selector}}, q, \{D_q^c, T_q\}\big)$
  $U_q \leftarrow (a_q, E_q)$
end if
// Repair: pattern-conditioned consolidation
$\hat{y} \leftarrow \mathbf{LLM}\big(\mathbf{P}_{\text{repair}}, q, U_q\big)$
return $\hat{y}$

MedExpQA (Alonso et al., 2024) contains medical licensing examination questions presented in multiple languages, together with reference explanations supporting their answers. We select its English test split, which contains 125 questions.

MedMCQA (Pal et al., 2022) is a comprehensive benchmark collected from Indian medical entrance examinations, covering 21 medical subjects and 2,400 healthcare topics. Since the labels of the official test set are not publicly available, we follow common practice and use the development set of 4,183 questions.

MedQA (Jin et al., 2021) is a medical benchmark derived from the United States Medical Licensing Examination (USMLE) and focuses on professional medical knowledge and clinical reasoning. We use its English test set, which contains 1,273 four-option questions.

MedXpertQA (Zuo et al., 2025) presents challenging clinical problems designed to test expert-level medical knowledge and reasoning. Our experiments are restricted to its 490 textual questions and exclude the multimodal portion of the benchmark.

MMLU-Med (Hendrycks et al., 2021) aggregates questions from six medical and biomedical MMLU subjects, including anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine. These subsets comprise 1,089 questions covering foundational biomedical and professional medical knowledge.

MedThink-Bench (Zhou et al., 2025) is a medical reasoning benchmark containing 500 challenging questions from ten medical domains. For each question, the dataset provides an expert-crafted reasoning trace and expert-defined evaluation points. We use the complete benchmark to evaluate both answer accuracy and the reliability of generated reasoning.

C.2 Evaluation Metrics.

We evaluate final-answer correctness on all eight benchmarks using **Accuracy**, defined as the percentage of examples for which the strictly parsed final option matches the gold answer.

Table 5: Dataset statistics for the eight medical QA benchmarks.

Dataset	Original size	Test size	# Options
MMLU-Pro	12,032	818	3–10
MedBullets	308	308	5
MedExpQA	2,488	125	4–5
MedMCQA	193,155	4,183	4
MedQA	61,097	1,273	4
MedXpertQA-Text	4,460	490	10
MMLU-Med	15,908	1,089	4
MedThink-Bench	500	500	3–22

On MedThink-Bench (Zhou et al., 2025), we further compare generated rationales with expert-annotated reasoning traces using four reference-based metrics. **METEOR** (Denkowski & Lavie, 2014) aligns unigram content through exact, stem, synonym, and paraphrase matches and combines precision and recall with a fragmentation penalty. **BERTScore F1** (Zhang et al., 2020) measures semantic similarity through contextual token embeddings and reports the harmonic mean of semantic precision and recall. **BLEU** (Chen & Cherry, 2014) is computed as macro-averaged, unsmoothed sentence-level BLEU-4 with a brevity penalty. **ROUGE-L F1** (Lin, 2004; Guo et al., 2024) measures ordered overlap using the longest common token subsequence between generated and reference rationales.

D Baseline Methods

To investigate the effectiveness of CARE, we compare it with representative approaches from two categories: inference methods without retrieval and retrieval-augmented methods.

D.1 Inference Methods without Retrieval.

These methods rely exclusively on the model's parametric knowledge and differ in the form of reasoning performed before answer generation.

Base directly predicts the answer using only the question and candidate options.

Chain-of-Thought (CoT) (Wei et al., 2022) prompts the model to produce intermediate reasoning steps before selecting the final answer.

CoT-Think (Yang et al., 2025) further activates the model's thinking mode, allowing extended internal reasoning without access to external evidence.

D.2 Retrieval-Augmented Methods.

These methods incorporate external medical evidence through different retrieval and reasoning strategies. To ensure a controlled comparison, all four RAG baselines use the same MedCorp3 Top-20 retrieval cache and Qwen3 inference interface. The three framework-based RAG baselines use the same settings as our experimental setup for fairness comparison.

Single-RAG (Lewis et al., 2020) follows a standard RAG pipeline. It combines the question and all answer options into a single query, retrieves 50 candidate documents with BM25, and uses MedCPT to rerank them. The five highest-ranked documents are then provided to the model for inference.

Adaptive-RAG (Jeong et al., 2024) dynamically selects a retrieval strategy according to the estimated complexity of each question. An LLM routes the question to no retrieval, single-round retrieval, or complex retrieval. For complex questions, the model produces a provisional answer, identifies the missing evidence, and expands the context with the top 10 documents from the retrieval cache.

TC-RAG (Jiang et al., 2025) formulates retrieval-augmented reasoning as a state-dependent process supported by stack-based memory. It divides the top 15 retrieved documents into three batches and

sequentially processes them with a planner, which uses PUSH, POP, and HALT operations to manage the evidence stack before generating the final answer from the retained documents.

MA-RAG (Wu et al., 2026) is a multi-round agentic RAG framework that uses disagreement among candidate responses to guide evidence refinement. In each round, eight agents sharing the same backbone model generate responses in parallel at a temperature of 0.7. Disagreement triggers dispute summarization and evidence expansion for the next round. The process runs for up to eight rounds, followed by majority voting if consensus is not reached.

E EXPERIMENTAL SETUP

E.1 MODEL AND GENERATION SETTINGS

We use Qwen3-4B, Qwen3-8B (Yang et al., 2025), and Qwen3.5-27B (Team, 2026) as the backbone models, all served through vLLM's OpenAI-compatible API. The maximum context length is set to 32,768 tokens, with prefix caching enabled. For Qwen3-4B and Qwen3-8B, the initial long-thinking stage generates up to 4,096 tokens, whereas Qwen3.5-27B is allowed up to 8,192 tokens. The Selector and final Repair stages are each limited to 1,000 generated tokens. All main results are obtained using a temperature of 0.0. The maximum concurrency is set to 64 for the 4B and 8B models and 32 for the 27B model. Because the number of answer choices varies across questions, the valid answer space is determined dynamically for each example.

E.2 RETRIEVAL DESIGN AND KNOWLEDGE SOURCES

We use MedCorp3 (Xiong et al., 2024) as the external knowledge source, including:

(1) **Textbooks**, a domain-specific collection of 18 widely used medical textbooks processed into snippets, providing foundational clinical principles and educational knowledge.

(2) **PubMed**, a comprehensive database containing 23.9 million biomedical abstracts, serving as a rich source of scientific evidence and contemporary research findings.

(3) **Wikipedia**, a large-scale general knowledge base comprising 6.5 million articles, providing high-level definitions and broad coverage of medical and general concepts.

Separate Lucene indexes are constructed offline for the three sources using Pyserini (Lin et al., 2021) and combined at the service layer as the `PubMedTextbooksWikipedia` collection. Retrieval follows a two-stage pipeline: BM25 (Robertson & Zaragoza, 2009) first retrieves a high-recall candidate set, and the MedCPT Cross-Encoder (Jin et al., 2023) subsequently reranks the candidates according to query-document relevance.

For initial retrieval, the question and all candidate options are combined into a single query. BM25 retrieves the top 50 documents, which are reranked to the top 20 by MedCPT. After deduplication based on document text, the five highest-ranked documents are provided to the initial long-thinking stage. For the Conflict pattern, a discriminative query is constructed from the first identified answer disagreement, followed by BM25 retrieval of the top 32 documents and MedCPT reranking to the top eight. For single-round-based baselines, we use the same initial retrieval settings, and for multi-round retrieval baselines, we also use the same settings with the conflict pattern.

F SUPPLEMENTARY RESULTS AND ANALYSIS

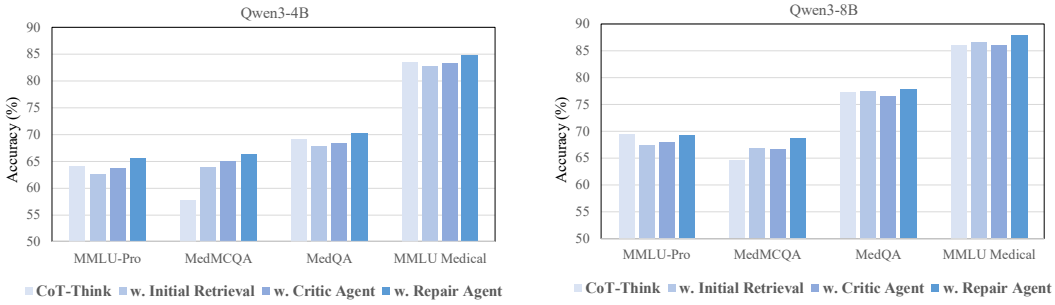
F.1 DETAILED REASONING RELIABILITY RESULTS

To complement the reasoning-reliability analysis in Section 4, we report the full results on MedThink-Bench in Table 6, corresponing the Figure 3. In addition to answer accuracy, we evaluate the agreement between model-generated and expert-annotated reasoning using METEOR, BERTScore F1, BLEU, and ROUGE-L F1.

As shown in Table 6, CARE achieves the best answer accuracy and reasoning-alignment scores on both backbones. Compared with MA-RAG, CARE improves accuracy from 34.40% to 39.20% on

Table 6: Detailed reasoning-reliability results on MedThink-Bench. We report answer accuracy and four metrics measuring agreement with expert-annotated reasoning. **Bold** indicates the best result for each backbone.

Model	Method	Accuracy	METEOR	BERTScore F1	BLEU	ROUGE-L F1
Qwen3-4B	CoT-Think	34.40%	18.49%	56.86%	1.28%	6.66%
	Single-RAG	36.60%	19.38%	56.66%	1.47%	6.82%
	Adaptive-RAG	36.40%	18.22%	55.89%	1.31%	6.23%
	TC-RAG	38.00%	13.70%	55.34%	0.91%	4.33%
	MA-RAG	34.40%	11.53%	57.26%	0.75%	3.63%
	CARE	**39.20%**	**30.69%**	**63.01%**	**6.54%**	**22.77%**
Qwen3-8B	CoT-Think	40.40%	17.55%	57.51%	1.26%	6.11%
	Single-RAG	41.60%	18.76%	57.06%	1.44%	6.45%
	Adaptive-RAG	41.20%	17.62%	56.22%	1.35%	5.94%
	TC-RAG	41.40%	13.56%	55.84%	0.93%	4.25%
	MA-RAG	42.00%	11.04%	57.74%	0.77%	3.43%
	CARE	**44.40%**	**32.02%**	**63.56%**	**6.82%**	**23.21%**



(a) Qwen3-4B. (b) Qwen3-8B.

Figure 11: Ablation study of CARE on Qwen3-4B and Qwen3-8B.

Qwen3-4B and from 42.00% to 44.40% on Qwen3-8B. More importantly, the improvements in reasoning alignment are substantially larger. On Qwen3-4B, CARE improves METEOR, BERTScore F1, BLEU, and ROUGE-L F1 over MA-RAG by 19.16, 5.75, 5.79, and 19.14 points, respectively. The corresponding improvements on Qwen3-8B are 20.98, 5.82, 6.05, and 19.78 points.

Averaged over the four reasoning-alignment metrics and the two backbone models, CARE improves upon MA-RAG by 12.81 points. These results further support the observation in Figure 3: improvements in final-answer accuracy alone do not fully characterize reasoning quality. CARE's trajectory-aware intervention produces reasoning that is substantially better aligned with expert annotations, in addition to improving the final answer.

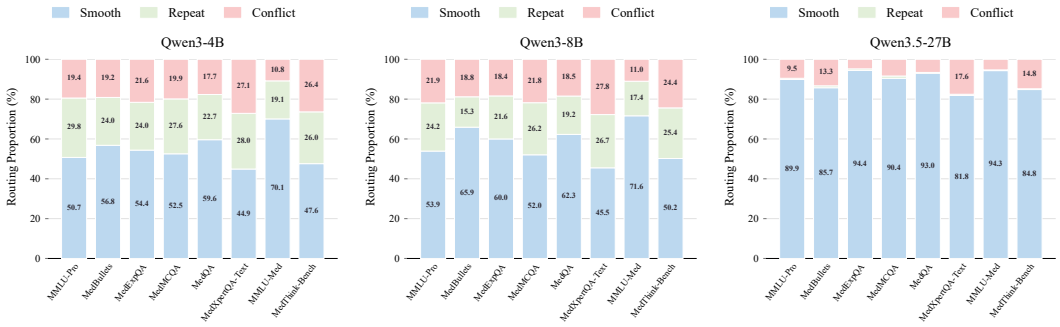
F.2 Ablation Study

We examine which components are responsible for CARE's improvements. Figure 11 presents the component-level ablations for both backbone models, while Table 7 studies the effect of evidence selection inside the Critic Agent.

The selector results show that the quality of evidence filtering has a direct effect on downstream reasoning. For Qwen3-4B, replacing the default selector with the distilled selector increases average accuracy from 58.25% to 59.28%, while the teacher selector reaches 62.54%. On Qwen3-8B, the corresponding results are 62.65%, 62.72%, and 65.59%. The teacher selector improves performance consistently across all eight benchmarks for both model sizes. These results suggest that retrieval alone is insufficient: after identifying a conflicting reasoning state, CARE also benefits from selecting evidence that is specifically relevant to the conflicting claims rather than indiscriminately injecting retrieved documents into the context.

Table 7: Analysis of different selectors used in the Critic Agent across different benchmarks.

Method	MMLU-P	MedB	MedE	MedMC	MedQA	MedX	MMLU-M	MedThink	Avg.
Backbone LLM: Qwen3-4B									
CoT-Think	64.06	50.65	68.80	57.78	69.05	14.49	82.64	34.40	55.23
CARE	65.65	48.70	72.00	66.36	70.31	18.98	84.76	39.20	58.25
w. Distill Selector	66.01	50.65	75.20	66.67	70.62	19.59	84.66	40.80	59.28
w. Teacher Selector	**69.68**	**57.14**	**76.00**	**69.45**	**75.96**	**21.63**	**86.87**	**43.60**	**62.54**
Backbone LLM: Qwen3-8B									
CoT-Think	69.56	57.79	73.60	64.55	77.22	16.33	86.41	40.40	60.73
CARE	69.32	56.82	74.40	68.78	77.77	21.84	87.88	44.40	62.65
w. Distill Selector	67.97	56.17	76.80	69.40	77.69	22.04	88.06	43.60	62.72
w. Teacher Selector	**72.13**	**60.39**	**76.80**	**71.91**	**81.62**	**25.51**	**89.99**	**46.40**	**65.59**



(a) Qwen3-4B. (b) Qwen3-8B. (c) Qwen3.5-27B.

Figure 12: Pattern distributions across datasets and model scales. Each bar shows the proportions of Smooth, Repeat, and Conflict patterns.

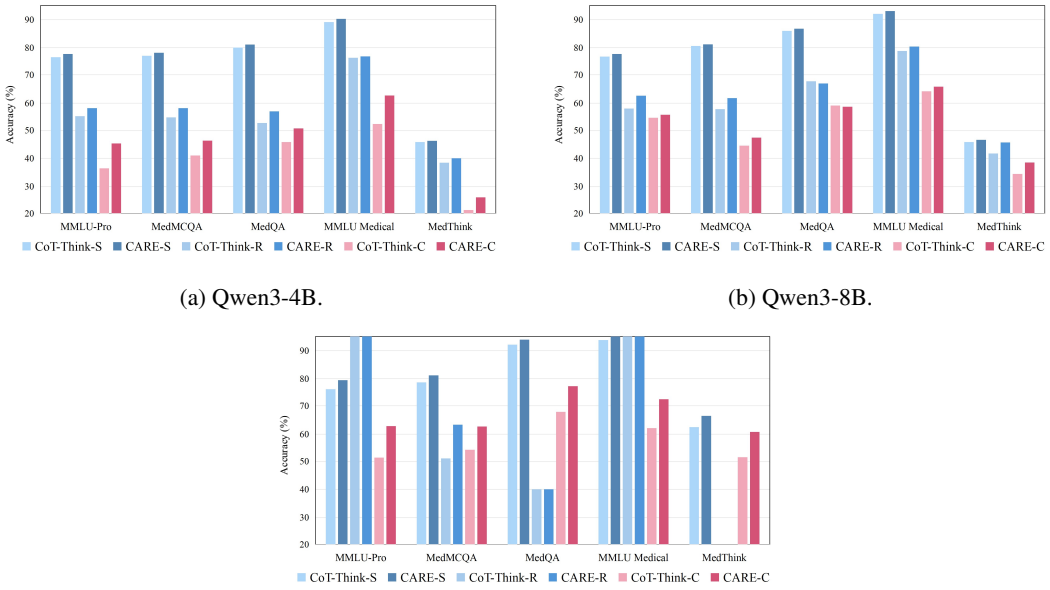
F.3 Pattern Analysis on the Full Dataset

We analyze the routing behavior of CARE across datasets and model scales to examine how reasoning stability relates to answer accuracy. The routing distributions are shown in Figure 12, while Figure 13 compares CoT-Think and CARE accuracy within each pattern. In Figure 13, "-S," "-R," and "-C" denote the Smooth, Repetitive, and Conflict branches, respectively.

From the experimental results, CARE improves the overall accuracy of all three models. Conflict questions generally have the lowest initial accuracy, confirming answer changes as a useful signal of reasoning difficulty. For Qwen3-4B, the Conflict branch improves from 38.64% to 43.56%. With increasing model capacity, the proportion assigned to the Smooth pattern reaches 90.32% for Qwen3.5-27B, yet CARE still improves its Conflict accuracy by 9.38 percentage points. Notably, Qwen3-8B shows slight degradation on the Repeat and Conflict branches of MedExpQA and MedQA, indicating that instability detection does not always yield beneficial evidence and that evidence selection remains a key source of error.

F.4 Analysis of Different Repeat Times k

We examine the effect of the repetition threshold k, defined as the number of consecutive identical answer commitments required to trigger trajectory truncation. As shown in Figure 14, accuracy improves substantially from $k = 1$ to $k = 2$ and remains relatively stable thereafter. Although the optimal value varies slightly across models and datasets, $k = 3$ consistently achieves near-optimal performance while avoiding premature truncation and unnecessarily prolonged reasoning. We therefore use $k = 3$ in all main experiments.

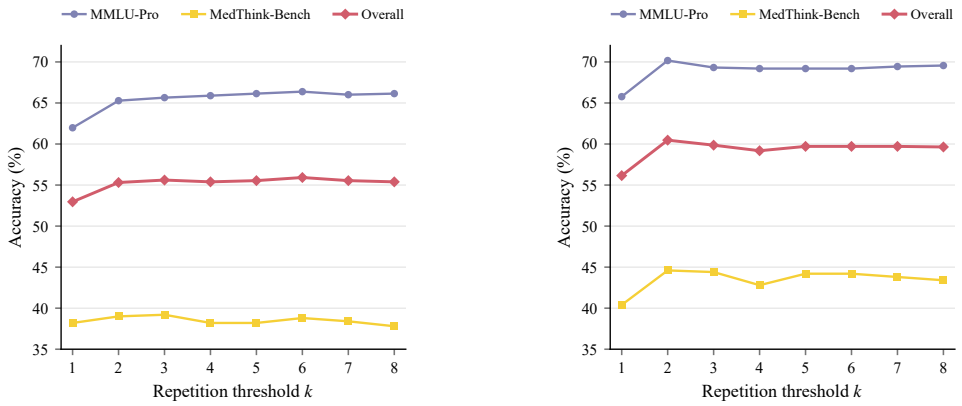


(a) Qwen3-4B.

(b) Qwen3-8B.

(c) Qwen3.5-27B.

Figure 13: Performance of models at different scales across different trajectory patterns. "-S" denotes the Smooth pattern, "-R" denotes the Repetitive pattern, and "-C" denotes the Conflict pattern.



(a) Qwen3-4B.

(b) Qwen3-8B.

Figure 14: Effect of answer accuracy to the repetition threshold k for Qwen3-4B and Qwen3-8B.

F.5 Retrieval for Repeat Patterns

We further investigate whether retrieval is beneficial for samples assigned to the *repeat* pattern. We compare standard chain-of-thought reasoning (*CoT-Think*), early stopping at the detected repeat point (*Repeat Early Stop*), and retrieval after repeat detection (*Repeat Retrieval*).

As shown in Table 8, early stopping at the repeat point already captures most of the improvement. Across both models, it increases overall accuracy from 55.59% to 58.36%, while additional retrieval yields only a marginal gain to 58.52%. Moreover, the effect of retrieval is inconsistent across datasets and can even degrade performance.

Table 8: Performance comparison of different strategies for the repeat pattern across benchmarks.

Method	MMLU-P	MedB	MedE	MedMC	MedQA	MedX	MMLU-M	MedThink	Avg.
Backbone LLM: Qwen3-4B									
CoT-Think	55.33	40.54	70.00	54.90	52.94	21.90	76.44	38.46	53.47
Repeat Early Stop	58.20	41.89	70.00	58.20	57.09	22.63	76.92	40.00	56.20
Repeat Retrieval	59.84	40.54	70.00	59.06	59.17	22.63	78.37	36.92	57.00
Backbone LLM: Qwen3-8B									
CoT-Think	58.08	55.32	70.37	57.86	67.76	23.66	78.84	41.73	57.92
Repeat Early Stop	62.63	59.57	66.67	61.79	66.94	22.90	80.42	45.67	60.74
Repeat Retrieval	62.12	61.70	70.37	61.33	66.53	25.95	77.78	41.73	60.20

These results suggest that once the model repeatedly converges to the same answer, additional external evidence provides limited benefit. This supports our design choice of retaining the truncated reasoning trajectory for the *repeat* pattern without invoking retrieval.

G Case Study

To further illustrate how CARE detects and resolves internal reasoning conflicts, we present a representative Conflict-pattern example from MedThink-Bench using Qwen3-4B under the self-critic configuration. The case shows how a correct final answer can conceal substantial instability within the reasoning trajectory and how the detected conflict guides targeted retrieval and final repair.

Case Study of Conflict-Guided Reasoning Repair

Question and Candidate Options.

A 4-week-old boy is brought to the pediatrician by his parents for an initial evaluation. His parents are concerned that he is not feeding well and has lost weight over the last 2 weeks. He was born at home via spontaneous vaginal delivery at 37 weeks of gestation to a G1P1 mother. His mother did not receive prenatal care. His blood pressure is 70/44 mmHg, pulse is 151/min, and respirations are 41/min. His weight is 3 kg (4th percentile), compared with 3.5 kg at birth (45th percentile). On physical examination, he is jaundiced with an enlarged liver and spleen. He does not appear to respond to visual stimuli, and further examination reveals bilateral clouding of the lens. Which of the following interventions could have avoided this patient's symptoms?

A. Avoiding fruit juice and sweetened foods
B. Changing to a soy-based formula
C. Providing imiglucerase enzyme replacement
D. Removing phenylalanine from the maternal diet during pregnancy
E. Vitamin B6 supplementation

Gold answer: B

Initial Long-Think: Diagnostic Direction.

> The key clues here are the jaundice, enlarged liver and spleen, cataracts, and the weight loss. ... The cataracts and liver issues are more in line with galactosemia.

Initial Long-Think: Unstable Option Comparison.

> The correct answer is probably C: imiglucerase. But that's for Gaucher disease. Wait, the symptoms here are jaundice, enlarged liver and spleen, cataracts, and weight loss. Those are more in line with galactosemia. So the correct intervention would be a galactose-free diet. But none of the options directly say that. However, option B is changing to a soy-based formula. Maybe that's the answer. But I'm not sure.

First Detected Conflict.
Previous commitment:

> Alternatively, maybe the answer is D: removing phenylalanine from the maternal diet during pregnancy. But that's for PKU, not galactosemia. The baby's symptoms don't match PKU.

New commitment:

> Alternatively, maybe the answer is C: imiglucerase. But that's for Gaucher disease. The symptoms here don't match that. The baby has cataracts, which is more in line with galactosemia.

Detected transition: $D \rightarrow C$

End of the Original Long-Think.

> If the baby was on a formula that had lactose, then changing to a soy formula would avoid galactose. But I'm not sure. Alternatively, maybe the answer is C. But I'm not sure. I think the answer is B.

Answer: B

Conflict-Guided Retrieval Query.
[Original question and candidate options]

> Medical evidence distinguishing these two candidate interpretations:
> - Removing phenylalanine from the maternal diet during pregnancy
> - Providing imiglucerase enzyme replacement

Selector Output.

> The patient presents with symptoms consistent with galactosemia, including jaundice, enlarged liver and spleen, cataracts, and poor weight gain. The key intervention to prevent these symptoms would be a galactose-free diet. Among the options, changing to a soy-based formula (B) is the most relevant, as it could potentially avoid galactose intake. Other options, such as imiglucerase (C), are for Gaucher disease; phenylalanine restriction (D) is for PKU; and the remaining options (A and E) are not directly related to the symptoms presented. Passage [4] discusses galactosemia and its symptoms, while passage [6] discusses vitamin B12 deficiency, which is not relevant here. Passage [0] is the only one that directly addresses the condition and intervention relevant to the question.

Final selection: [4] [6] [0] [2] [1]

Final Repair Output.
Reasoning:

> The patient's symptoms (jaundice, enlarged liver and spleen, cataracts, and poor weight gain) are consistent with galactosemia, a metabolic disorder in which the body cannot process galactose. A galactose-free diet is the primary intervention. While none of the options directly states this, changing to a soy-based formula (B) may help avoid galactose intake, although soy formula is not inherently galactose-free. The other options (A, C, D, and E) are not relevant to galactosemia.

Answer choice: B

In this case, the patient's feeding difficulty, jaundice, hepatosplenomegaly, poor weight gain, and bilateral cataracts strongly indicate classic galactosemia. Although the model recognizes this diagnosis early, uncertainty about the soy-formula option causes repeated shifts among B, C, and D.

The trajectory eventually reaches the correct answer but contains 16 answer commitments, revealing substantial instability hidden by answer accuracy alone.

CARE detects the first transition from D to C and constructs a discriminative query contrasting dietary intervention for PKU with imiglucerase treatment for Gaucher disease. BM25 retrieves 32 candidates, and MedCPT reranks the top eight. The Selector then compares them with the original long-thinking trajectory `[0]` and retains `[4, 6, 0, 2, 1]`. It correctly links C to Gaucher disease, D to PKU, and the symptoms to galactosemia, supporting B as the most appropriate answer.

This case also exposes an evidence-attribution error. The Selector claims that passage `[4]` discusses galactosemia, although it concerns childhood development and autism, while passage `[6]` addresses infant vitamin B12 deficiency and offers limited diagnostic value. This shows that sharing the same backbone across Solver and Critic roles can lead to inaccurate summaries of external evidence. The Repair Agent therefore treats both the Selector analysis and retrieved documents as provisional inputs and independently verifies their medical claims, reducing the risk that incorrect evidence attribution propagates into a new hallucination.

H PROMPT DETAILS

This section presents the system instructions and prompt templates used by the Solver, Critic, Selector, and Repair agents.

System Prompts

 # Solver / refine / decision
You are a careful medical multiple-choice reasoning assistant.

 # Selector
You select relevant medical passages.

Solver Agent Prompt

Solve this medical multiple-choice question carefully.

The following retrieved passages may be useful, irrelevant, incomplete, or misleading. Evaluate them against the question rather than assuming they are correct or sufficient.

 # Retrieved Documents
[1] {title_1}
{content_1}
...
[5] {title_5}
{content_5}

 # Question
{question_and_options}

Reason independently. {answer_rule} End with exactly one final answer tag, where every letter is one of: {valid_letters}.
/think

Repair Agent Prompt for Smooth and Repeat Patterns

Review and reorganize the draft reasoning for this medical question. Treat the draft as fallible working notes, not as authoritative evidence. Independently verify its medical claims against the question and options, correct any mistake or unsupported conclusion, and produce a concise final rationale.

Return JSON only as Dict{"reasoning": Str(concise verified explanation), "answer_choice": Str{A/B/...}}.

 # Question
{question_and_options}

 # Draft reasoning
{full_or_truncated_longthink}

Return JSON only as Dict{"reasoning": Str(concise verified explanation), "answer_choice": Str{A/B/...}}.
/no_think

Critic Prompt for Conflict-Guided Retrieval

{raw_question_and_options}
Medical evidence distinguishing these two candidate interpretations:
- {text_of_old_answer_option}
- {text_of_new_answer_option}

Critic Prompt for Passage Selection

I will provide you with 9 passages, each indicated by a numerical identifier [id]. Your task is to select the top-5 most useful passages for answering the given medical question:
{question_and_options}

 # Documents
[0]
{original_long_think}
[1]
{reranked_document_1}
...
[8]
{reranked_document_8}

 # Question
{question_and_options}

Please follow the steps carefully:

Step 1. Information Requirements Identification:
Please list up the key knowledge points to answer the question thoroughly.

Step 2. Requirement-to-Passage Mapping:
Map passages to each knowledge point in Step 1, group them accordingly, and put irrelevant ones into a "No Useful Information" group.

Step 3. Document Selection for Completeness and Conciseness:
Select up to 5 passages that maximize coverage across key knowledge points while avoiding redundancy.

Your output should follow the following format, do not output other additional content:
- Reasoning: [Detailed explanation]
- Final Selection: [id1] [id2] [id3] [id4] [id5]
/no_think

Repair Agent Prompt for the Conflict Pattern

You are a professional medical expert to answer the # Question using the # Retrieved Documents. Please answer the question carefully and give a definite answer.
First extract and utilize useful information from the retrieved documents. If they are insufficient or irrelevant, rely on your own medical knowledge. The Selector Analysis is advisory reasoning about information needs and document relevance. Check it against the selected documents and the question; do not treat unsupported claims in it as medical evidence.
Return JSON only as Dict{"reasoning": Str(concise explanation), "answer_choice": Str{A/B/...}}.

 # Selector Analysis
{selector_analysis_without_final_selection}

 # Retrieved Documents
{selected_complete_units_in_selector_order}

 # Question
{question_and_options}

Return JSON only as Dict{"reasoning": Str(concise explanation), "answer_choice": Str{A/B/...}}.
/no_think