南京大学

From Conflict to Consensus

Boosting Medical Reasoning via Multi-Round Agentic RAG

Wenhao Wu, Zhentao Tang, Yafu Li, Shixiong Kai, Mingxuan Yuan, Zhenhong Sun, Chunlin Chen, Zhi Wang

Nanjing University, Huawei Noah's Ark Lab, The Chinese University of Hong Kong, Australian National University

Email: wenhaowu@smail.nju.edu.cn

Code: https://github.com/NJU-RL/MA-RAG

Contents

1 Background: Test-Time Scaling

➢ Parallel scaling

- **parallel exploration**: generate multiple independent outputs for a given prompt
- lack **coordination** across samples and cannot support **iterative refinement**
- output selection:
  - unsupervised methods (e.g., majority voting) to pick the most consistent result
  - external verifiers for evaluation (Best-of-N)

➢ Sequential scaling

- **long chain-of-thought reasoning** (verification, reflection, backtracking, subgoal decomposition) before outputting the final answer
- multi-round feedback for iterative **refinement**
- **error accumulation**: easily stuck in incorrect reasoning paths and struggle to recover correct conclusions

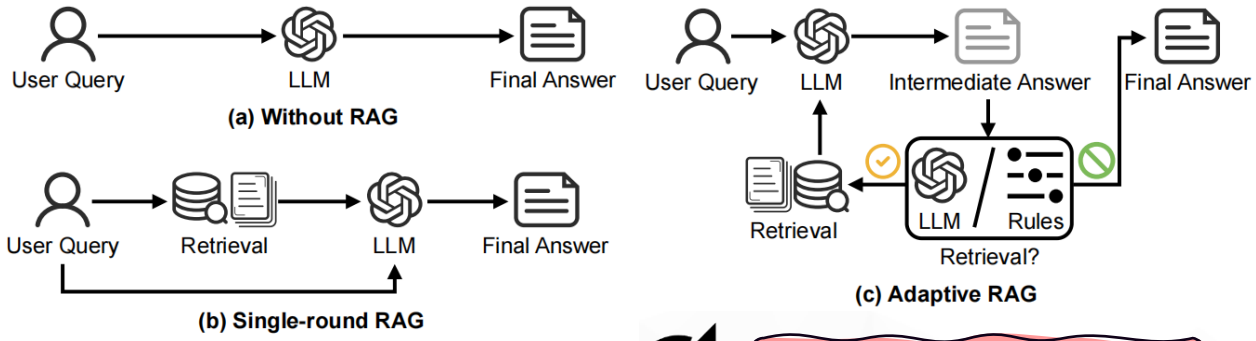
➢ Hybrid scaling

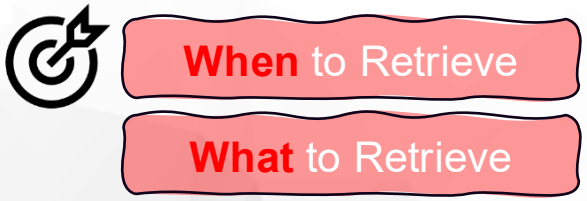
- integrate strengths of **parallel exploration** and **sequential iterative refinement**
- benefit from parallel exploration, perform self-verification and iterative refinement

1 Background: RAG

➢ Application risks of LLMs

- **Outdated Knowledge**: The knowledge in the model parameters is static and cannot be updated in a timely manner.
- **Hallucination**: The model may generate content that seems reasonable but is inconsistent with facts.
- **Safety Concern**: Healthcare is a high-risk field, which prioritizes interpretable, evidence-based responses.



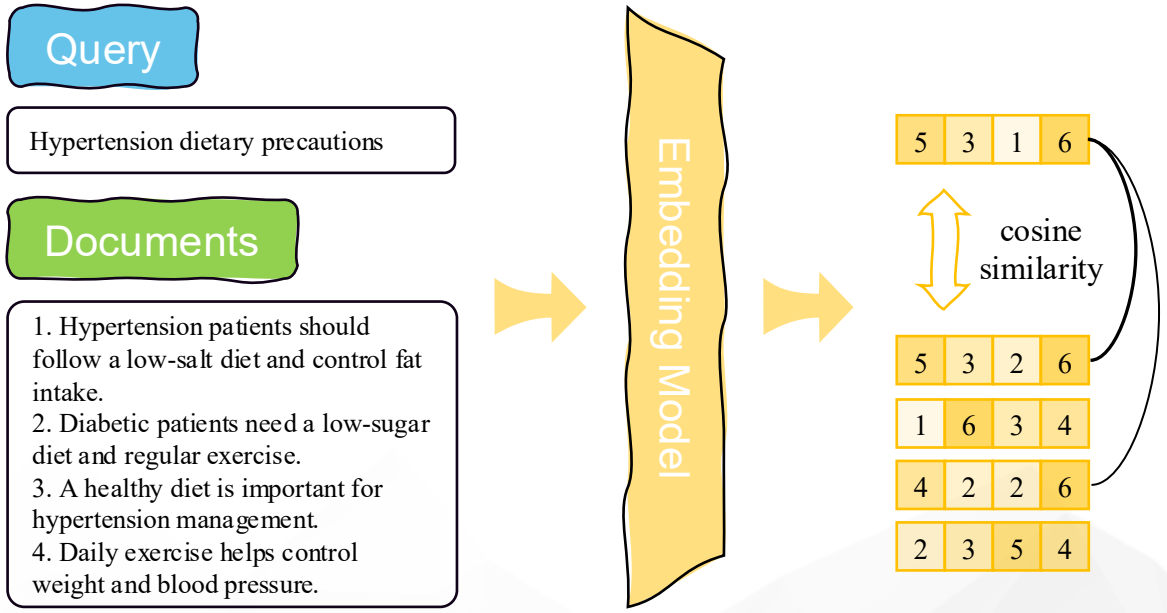


➢ Retrieval-Augmented Generation (RAG)

- **Sparse retrieval - BM25**

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}$$

- **Dense retrieval - embedding**



1 Background: Limitations

➢ Over confidence & Noisy token-level metrics

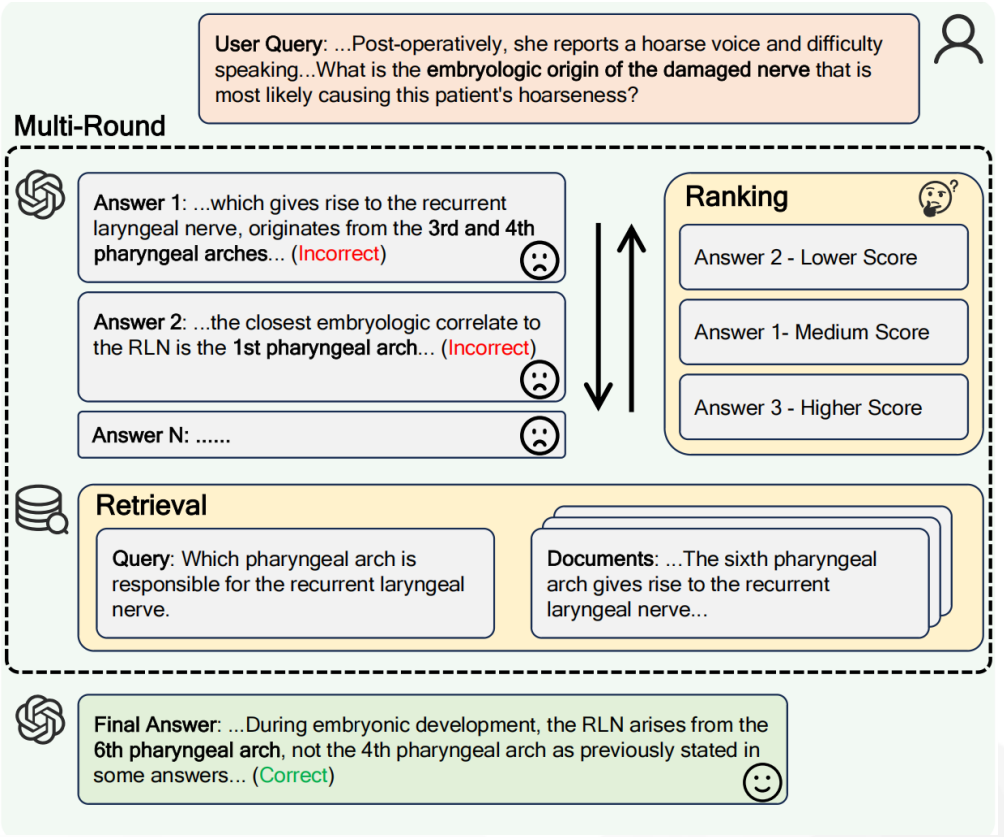
- LLMs frequently generate **hallucinations with high confidence**, making token-level uncertainty a weak indicator of actual retrieval needs
- Uncertainty estimates at the token level are often dominated by trivial words, rather than domain-critical medical concepts, leading to imprecise query formulation

➢ Irrelevant information degrades performance

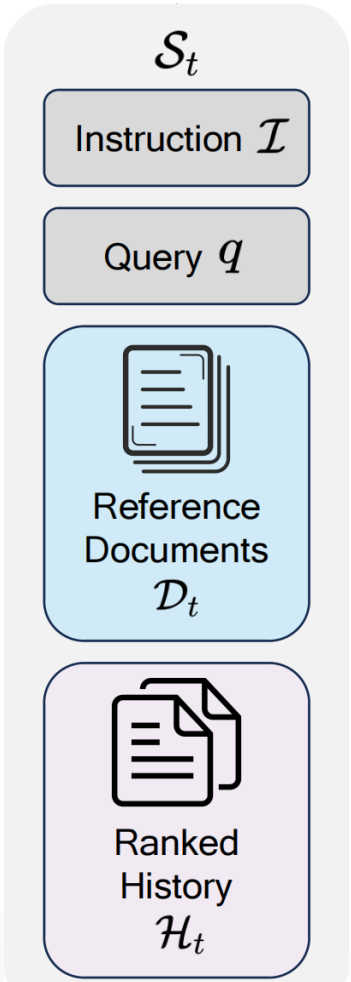
- Retrieval can introduce **irrelevant noise** into the context, which directly degrades model performance when retrieval is not universally beneficial
- The model cannot effectively capture and utilize core evidence, further impairing the performance of complex medical reasoning

Could we bypass the reliance on noisy token-level signals by leveraging higher-level semantic cues to steer agentic retrieval more efficiently?

➢ Inconsistency means a lack of knowledge



2 Method: Problem Statement



➢ **Iterative context optimization during multi-round agentic refinement**

$$S_t = \{I, q, \mathcal{D}_t, \mathcal{H}_t\}, \qquad t \in \{1, \dots, T\}$$

- I: task instruction
- q: user query

} **invariant**

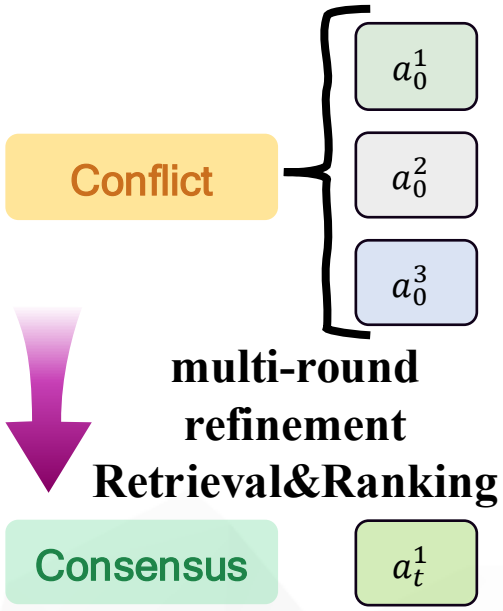
- $\mathcal{D}_t$: document context, retrieved documents by Retrieval Agent
- $\mathcal{H}_t = \text{Rank}(\mathcal{A}_{t-1})$: history context, ranked previous answers by Ranking Agent

} **evolvable**

How to effectively transition state S_t to the new round S_{t+1}?

2 Method: Overview

- **MA-RAG**
  - Solver Agent
  - Retrieval Agent
  - Ranking Agent



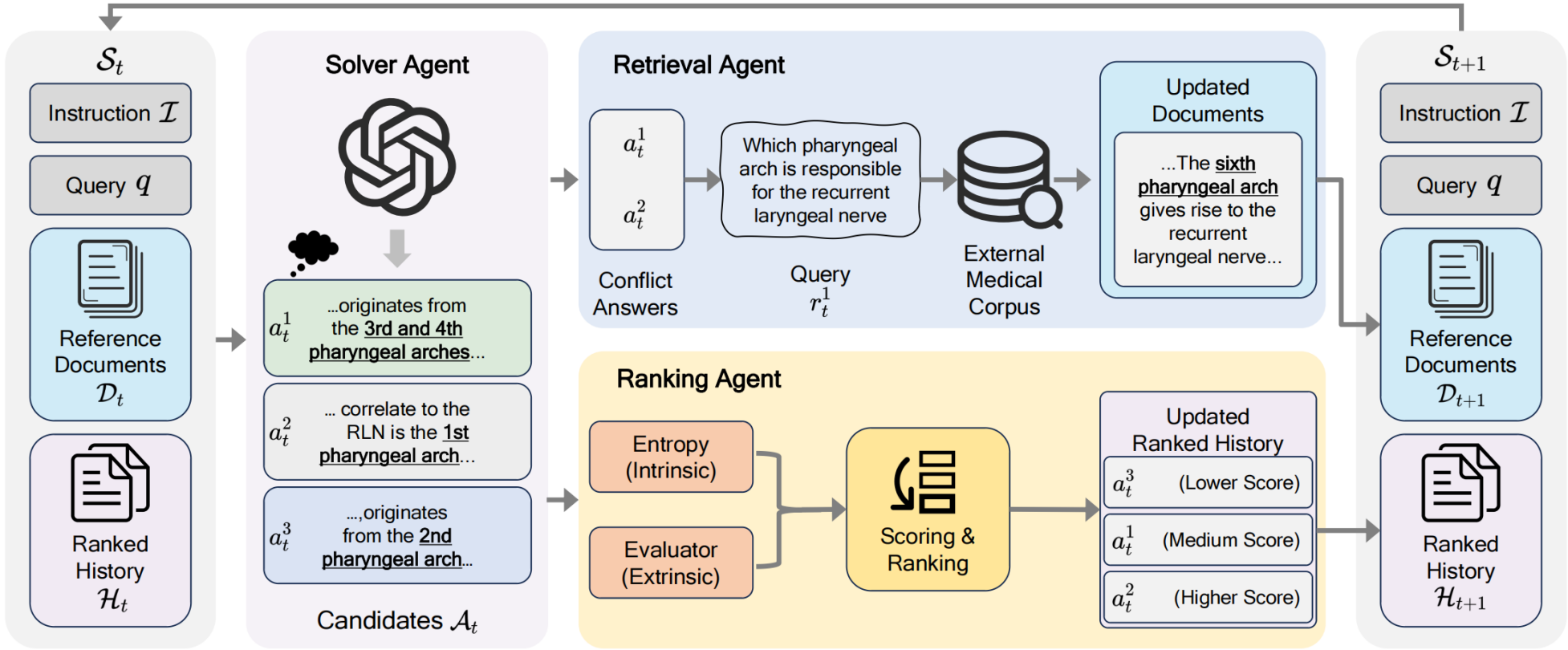
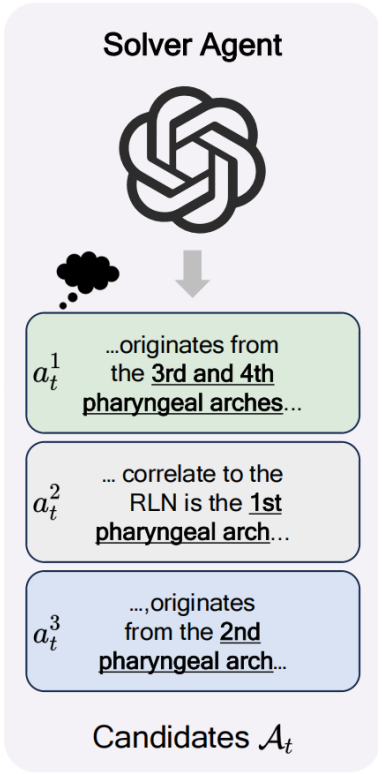


Figure 2. Overall pipeline of **MA-RAG** (**M**ulti-round **A**gentic RAG) for complex medical reasoning. At each round t of the agentic refinement loop: i) the **Solver Agent** first samples a diverse set of candidate responses; ii) the **Retrieval Agent** transforms semantic conflicts among candidates into actionable queries to retrieve external evidence from a local medical corpus, updating the document context to $\mathcal{D}_{t+1}$; and iii) the **Ranking Agent** restructures history reasoning traces $\mathcal{A}_t$ by prioritizing top-tier candidates to construct the history context $\mathcal{H}_{t+1}$, mitigating long-context degradation and enhancing in-context learning. The evolved state $S_t = \{\mathcal{I}, q, \mathcal{D}_t, \mathcal{H}_t\}$ serves as the prompt at the next round, iteratively rectifying semantic **conflict** toward converging to a reliable, high-fidelity **consensus**.

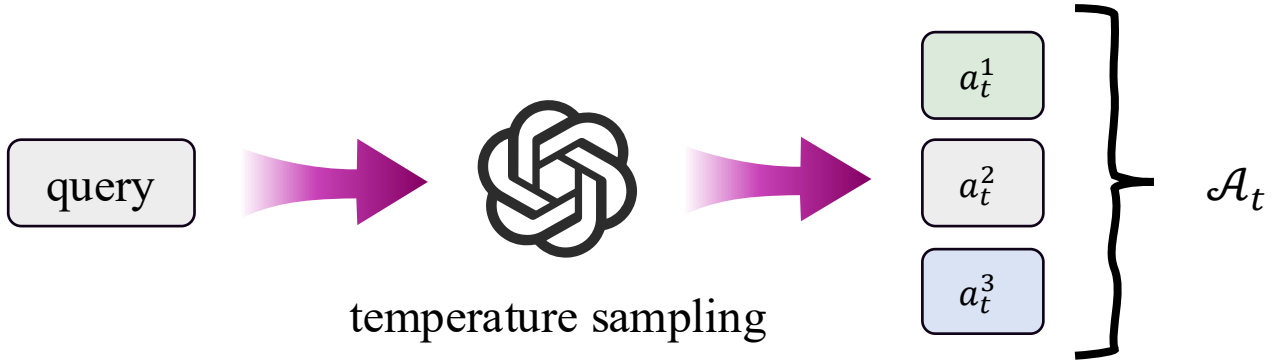
2 Method: Solver Agent



- sample N candidate responses $\mathcal{A}_t$
- exploring diverse reasoning paths

➢ **Diversity and exploration**

$$\mathcal{A}_t = \{a_t^1, a_t^2, \ldots, a_t^N\} \sim \mathcal{M}(I_{\text{solver}}, q, \mathcal{D}_t, \mathcal{H}_t)$$



➢ **Transform static self-consistency into an adaptive scaling mechanism**

accurate reasoning chains tend to converge toward a stable consensus → Final answer

hallucinations often exhibit divergent inconsistencies → How to utilize these conflicts to derive a consensus

2 Method: Retrieval Agent

➢ **When and What**

- when to trigger retrieve
- use what to retrieve

➢ **Token-level Uncertainty (Unreliable)**

- probabilities, entropies, attention weights…
- tokens with low probability/high entropy

LLMs tend to generate hallucinations with over-confidence

Answer 1: ...which gives rise to the recurrent laryngeal nerve, originates from the 3rd and 4th pharyngeal arches... (Incorrect)

Answer 2: ...the closest embryologic correlate to the RLN is the 1st pharyngeal arch... (Incorrect)

Answer N: .....

➢ **Semantic-level Conflicts (Simple but Effective)**

- sufficient knowledge and reasoning capacity lead to **self-consistency** across multiple independent generations
- semantic conflicts indicate the model's current knowledge **deficiency**
- Retrieval Agent extracts candidate conflicts and formulates targeted retrieval queries $\mathcal{R}_t$

When to retrieve?

Conflicts

What to retrieve?

actionable queries

$$\mathcal{R}_t = \{r_t^1, r_t^2, \ldots, r_t^K\} \sim \mathcal{M}(I_{\text{conflict}}, q, \mathcal{A}_t)$$

r_t^1 r_t^2 r_t^3

update document context

$\mathcal{D}_t \rightarrow \mathcal{D}_{t+1}$

medical corpus

➢ **Semantic conflict as a reliable indicator**

2 Method: Ranking Agent

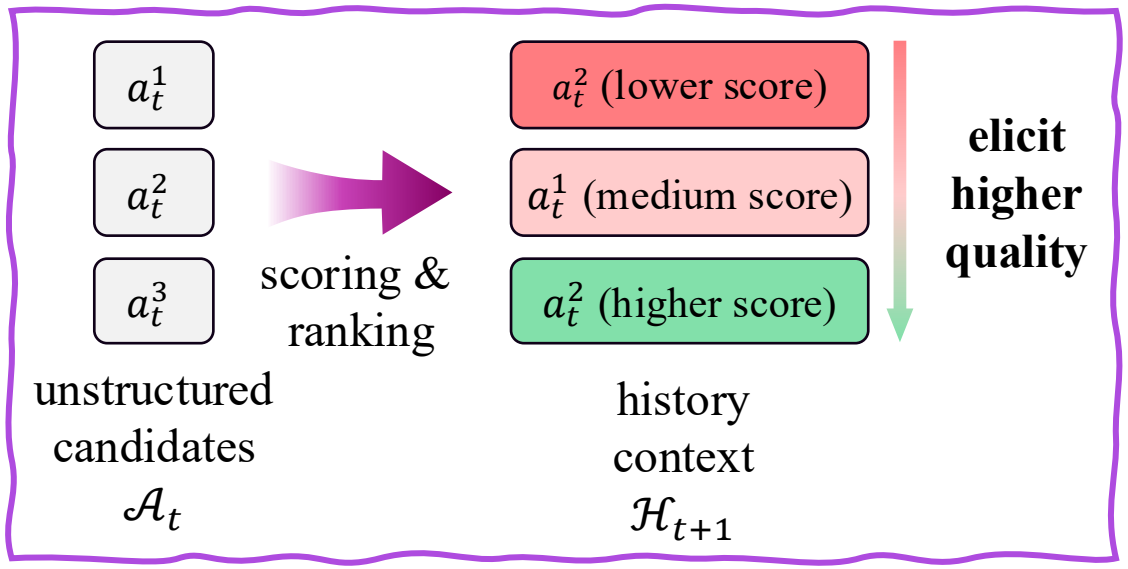
lost-in-the-middle → models may overlook critical reasoning cues **situated centrally** within an expanding prompt

How to optimize history context to enhance in-context learning?

➢ Restructure History Context

$$\mathcal{H}_t = \text{sort}(\mathcal{A}_{t-1}, \text{key} = Q) = \left(a_{t-1}^{(1)}, a_{t-1}^{(2)}, \ldots, a_{t-1}^{(N)}\right)$$

- evaluate the quality of candidate responses in the previous round $\mathcal{A}_{t-1}$, and constructs the history context $\mathcal{H}_t$
- scoring function Q: **Intrinsic Uncertainty**, **Extrinsic Verification**



➢ Intrinsic Uncertainty: sequence-level entropy

- simple, cheap metric for estimating quality (**general**)

$$Q_{\text{int}}(a) = -\frac{1}{L}\sum_{i=1}^{L} \text{entropy}\left(P(x_i | x_{<i}, S_t)\right)$$

➢ Extrinsic Verification: fine-tuned evaluator

- fine-tuned binary classifier V_θ (**semantic correctness**)
- datasets: question and sample diverse responses $\mathcal{D}_{qa} = \{(q_i, a_i)\}_{i=1}^{M}$

$$Q_{\text{ext}} = \text{sigmoid}\left(V_\theta(q, a)\right)$$

3 Experiments: Main Results

- **Backbones**
  - Qwen3-8B
  - Llama-3.1-8B-Instruct
  - UltraMedical-3.1-8B
  - HuatuoGPT-o1-8B
- **Test-time scaling**
  - Chain-of-Thought
  - Self-Consistent
  - MDAgents
  - Multi-Refine
- **Naïve RAG**
  - Single-round RAG
  - Fix-length RAG
  - Fix-sentence RAG
- **Adaptive RAG**
  - FLARE
  - TC-RAG

Table 1. Main results (Accuracy %) on seven medical benchmarks. The best results are highlighted in **bold**, and the second-best are underlined. MA-RAG-int/MA-RAG-ext denote our method using the intrinsic uncertainty/extrinsic verification in the ranking agent.

Method	MedQA	MedMCQA	Medbullets	MMLU-Pro	NEJM	MedExpQA	MedXpertQA	Avg.
General and Medical LLMs								
Qwen3-8B	71.1	61.3	51.0	64.9	56.0	67.2	16.1	55.4
Llama-3.1-8B	68.1	58.1	47.1	58.0	50.5	67.2	12.4	51.6
UltraMedical-3.1-8B	69.6	56.5	54.5	48.5	41.5	66.4	13.1	50.0
HuatuoGPT-o1-8B	71.1	60.2	50.6	54.0	47.9	63.2	15.5	51.8
Test-Time Scaling Methods (Backbone: Qwen3-8B)								
CoT	69.3	60.3	50.6	63.4	56.2	72.0	18.0	55.7
SC	73.3	62.4	51.9	66.5	56.6	70.4	15.7	56.7
MDAgents	72.1	**67.5**	52.2	65.7	57.1	74.4	18.2	58.2
Multi-Refine	74.7	63.6	55.8	69.1	58.0	73.6	16.3	58.7
RAG Methods (Backbone: Qwen3-8B)								
SR-RAG	69.9	64.4	49.4	66.1	57.6	72.8	17.3	56.8
FL-RAG	69.6	63.3	52.3	64.1	57.2	69.6	17.0	56.2
FS-RAG	68.0	61.3	51.3	59.5	53.1	67.2	16.5	53.8
FLARE	72.7	61.7	51.9	62.2	55.3	71.2	17.7	56.1
TC-RAG	70.0	64.6	49.0	63.6	57.7	75.2	18.0	56.9
MA-RAG-int	77.0	67.1	57.1	**70.9**	**60.8**	72.8	21.2	61.0
MA-RAG-ext	**77.1**	67.2	**59.1**	70.7	60.5	**78.4**	**22.2**	**62.2**

Superiority over test-time scaling & RAG baselines

3 Experiments: Ablations & Scalability

Table 2. Ablation study on the components of MA-RAG using Qwen3-8B backbone. The best results are highlighted in **bold**.

Method	MedQA	MedMCQA	Medbullets	MMLU-Pro	NEJM	MedExpQA	MedXpertQA	Avg.
Qwen3-8B	71.1	61.3	51.0	64.9	56.0	67.2	16.1	55.4
+Multi-Refine	74.7	63.6	55.8	69.1	58.0	73.6	16.3	58.7
+Retrieval Agent	**77.1**	66.2	57.5	69.6	60.3	74.4	19.2	60.6
+Ranking Agent	**77.1**	**67.2**	**59.1**	**70.7**	**60.5**	**78.4**	**22.2**	**62.2**

Table 4. Performance of MA-RAG's ablations based on the Qwen3-32B backbone, demonstrating its scalability across model capacities.

Method	MedQA	MedMCQA	Medbullets	MMLU-Pro	NEJM	MedExpQA	MedXpertQA	Avg.
Qwen3-32B	80.8	68.8	63.0	73.1	64.4	82.4	20.2	64.7
+Multi-Refine	83.6	70.9	64.9	75.5	68.4	86.4	21.2	67.3
+Retrieval Agent	83.5	72.8	70.8	76.3	69.2	86.4	24.9	69.1
+Ranking Agent	**85.3**	**73.4**	**71.4**	**76.7**	**70.4**	**87.2**	**27.3**	**70.2**

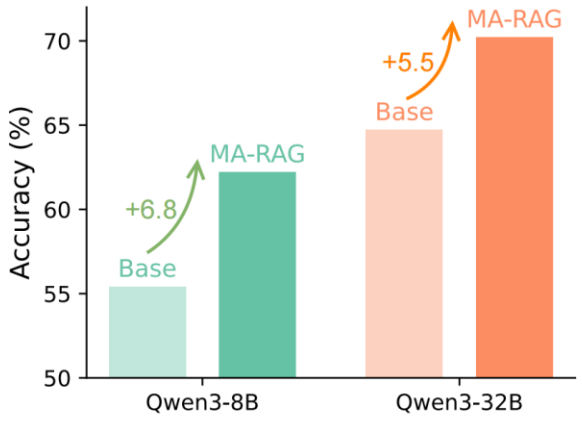


Figure 5. Performance of MA-RAG across backbone model scales.

Efficacy of the Retrieval & Ranking Agent

Base → +Multi-Refine → +Retrieval Agent → +Ranking Agent

3 Experiments: Ranking Agent Effect

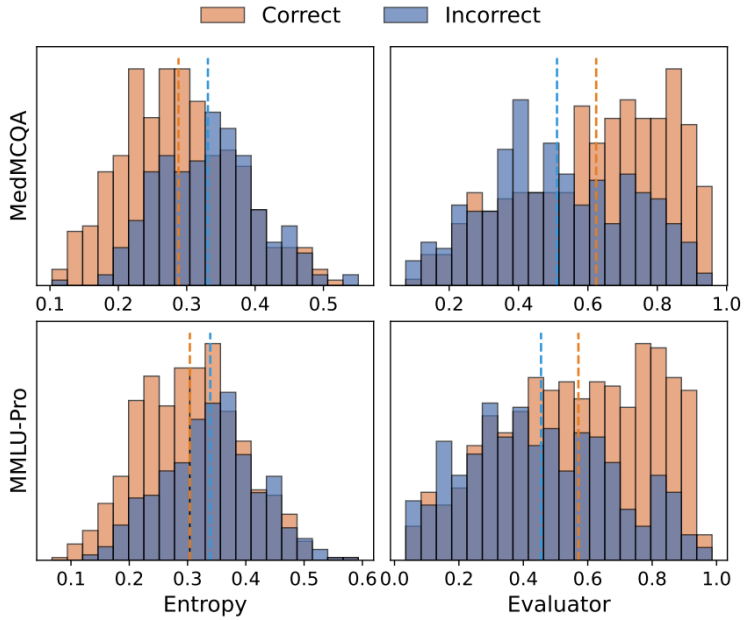


Figure 4. Visualization of the response score density on MedMCQA and MMLU-Pro. **Intrinsic Uncertainty** (left): Correct answers exhibit lower entropy compared to incorrect ones. **Extrinsic Verification** (right): The fine-tuned BERT-based evaluator assigns higher scores to correct answers, exhibiting a more pronounced discriminative margin than the entropy-based score. These distinct distributions validate the utility of both score functions for the Ranking Agent. Notably, the extrinsic evaluator exhibits superior discriminative power compared to the intrinsic counterpart, a finding consistent with the performance gap observed between MA-RAG-int and MA-RAG-ext in Table 1.

Effectiveness of entropy & evaluator

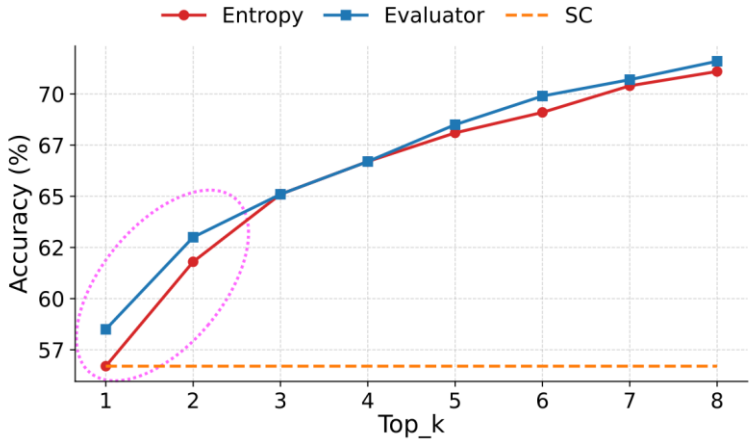


Figure 8. Recall@k performance curves for Intrinsic Entropy and Extrinsic Evaluator rankings as k increases from 1 to 8. The circled regions at $k=1$ and $k=2$ highlight the steep improvement in recall, demonstrating that while the top-1 candidate often misses the ground truth, the correct answer is frequently covered within the top-2. This sharp rise validates MA-RAG's design, which leverages the broader context of top-ranked candidates rather than relying on a single static selection.

Table 13. Comparison between static Best-of-N re-ranking and MA-RAG under different scoring mechanisms on Qwen3-8B. We report `Recall@k` to measure whether the correct answer appears in the top-k ranked candidates scored by Intrinsic Entropy or an Extrinsic Evaluator, and include `Pass@20` as the oracle upper bound of the candidate pool.

Method	MedQA	MedMCQA	Medbullets	MMLU-Pro	NEJM	MedExpQA	MedXpertQA	Avg.
SC	73.3	62.4	51.9	66.5	56.6	70.4	15.7	56.7
Pass@20	90.1	83.7	78.2	82.6	77.2	90.4	40.2	77.5
Intrinsic Entropy								
Recall@1	73.1	61.9	52.6	66.4	55.7	71.2	16.3	56.7
Recall@2	78.3	67.5	60.4	71.3	62.0	72.8	20.6	61.8
MA-RAG-int	77.0	67.1	57.1	70.9	60.8	72.8	21.2	61.0
Extrinsic Evaluator								
Recall@1	73.7	62.0	56.8	66.3	53.2	76.8	20.4	58.5
Recall@2	77.6	67.7	62.3	71.6	59.8	76.8	25.3	63.0
MA-RAG-ext	77.1	67.2	59.1	70.7	60.5	78.4	22.2	62.2

Superiority over Best-of-N

3 Experiments: Test-Time Scaling Analysis

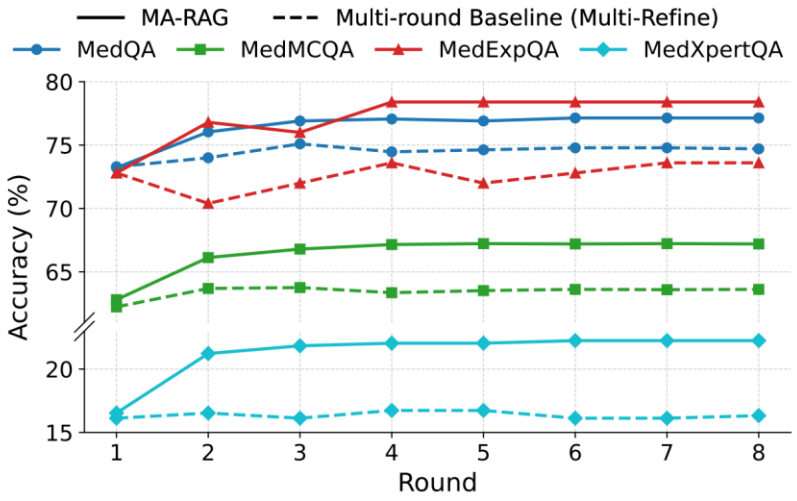


Figure 3. Performance comparison between MA-RAG and the multi-round test-time scaling baseline (Multi-Refine).

Much more stable during multi-round refinement

Table 3. MA-RAG's test-time scaling performance w.r.t. the maximum refinement rounds T and the size of the candidate pool N.

MA-RAG-ext	MedQA	MedMCQA	Medbullets	MMLU-Pro	NEJM	MedExpQA	MedXpertQA	Avg.
Maximum Refinement Rounds T								
T = 1	73.2	62.8	53.2	67.6	56.2	72.8	16.5	57.5
T = 2	76.0	66.1	**60.4**	69.8	60.2	76.8	21.2	61.5
T = 4	**77.1**	**67.2**	59.1	**70.8**	60.2	**78.4**	22.0	62.1
T = 8	**77.1**	**67.2**	59.1	70.7	**60.5**	**78.4**	**22.2**	**62.2**
Size of Candidate Pool N								
N = 2	74.5	64.7	54.9	69.8	56.6	72.0	17.3	58.5
N = 4	76.0	65.7	58.4	68.2	58.8	74.4	18.6	60.0
N = 8	**77.1**	**67.2**	**59.1**	**70.7**	**60.5**	**78.4**	**22.2**	**62.2**

Table 9. Performance of MA-RAG-ext under varying query granularities (K) using the Qwen3-8B backbone. We vary the number of conflict-driven queries generated per round while maintaining a fixed global retrieval budget of $|\mathcal{D}| = 8$ documents. The best results are highlighted in **bold**.

K	MedQA	MedMCQA	Medbullets	MMLU-Pro	NEJM	MedExpQA	MedXpertQA	Avg.
$K=1$	75.3	**67.7**	60.1	70.4	58.9	76.8	20.8	61.4
$K=2$	**77.3**	67.3	**60.3**	**70.8**	58.5	76.8	21.2	61.7
$K=4$	77.1	67.2	59.1	70.7	**60.5**	**78.4**	**22.2**	**62.2**

Test-time scaling properties across size of rounds, candidates, and conflict-aware queries

3 Experiments: Inference Efficiency

Table 5. Inference efficiency comparison across methods on MedXpertQA with Qwen3-8B. We report average retrieved documents, generated tokens, final answer tokens, and wall-clock time per question.

Method	Avg. Docs	Avg. Gen. Tokens	Avg. Final Tokens	Avg. Time (s)	Accuracy
Base (Qwen3-8B)	0	514	514	5.6	16.1
Self-Consistency (SC)	0	10,095	505	28.0	15.7
FLARE	30.0	962	429	42.7	17.7
TC-RAG	11.5	1,260	1,067	44.5	18.0
MDAgents	8.0	7,546	–	146.2	18.2
MA-RAG-ext	13.7	12,762	589	70.7	**22.2**

Table 17. Inference efficiency comparison across methods on Medbullets with Qwen3-8B. We report average retrieved documents, generated tokens, final answer tokens, and wall-clock time per question.

Method	Avg. Docs	Avg. Gen. Tokens	Avg. Final Tokens	Avg. Time (s)	Accuracy
Base (Qwen3-8B)	0	449	449	5.6	51.0
Self-Consistency (SC)	0	9,016	451	22.0	51.9
FLARE	32.2	767	377	32.2	51.9
TC-RAG	12.9	1,229	1,177	49.0	49.0
MDAgents	8.0	5,810	–	115.4	52.2
MA-RAG-ext	9.0	8,810	496	41.1	**59.1**

More grounded and reliable retrieval signal
Favorable cost-performance trade-off

4 Conclusions

- **Innovative Framework**: This study proposes MA-RAG (Multi-Round Agentic RAG), an innovative multi-round agentic refinement framework that iteratively evolves external evidence and internal reasoning history to boost test-time scaling for complex medical reasoning.

- **Key Contributions**: MA-RAG comprises three core agents (Solver, Retrieval, Ranking) for candidate generation, evidence retrieval and reasoning history optimization, extending the self-consistency principle and integrating a boosting mechanism to minimize residual reasoning errors.

- **Significant Advantages**: MA-RAG delivers substantial **+6.8** average accuracy over the backbone model across 7 medical Q&A benchmarks, and outperforms test-time scaling and RAG baselines. It shows exceptional performance on **complex expert-level** medical reasoning tasks and strong scalability across different model sizes.

- **Future Prospects**: Future work would reduce the inference-time cost for real-world clinical deployment, integrate richer medical resources (web-scale search, structured medical databases) to broaden evidence coverage, and develop more robust answer quality evaluation metrics to further unlock the potential of iterative reasoning optimization.

南京大学

Thank you!

Wenhao Wu

Nanjing University

Email: wenhaowu@smail.nju.edu.cn

Code: https://github.com/NJU-RL/MA-RAG

Paper: https://arxiv.org/abs/2603.03292