

MERMAIDS: Structured Memory and Evidence Reuse for Reducing Multi-Hop Retrieval Hallucinations in Agentic RAG

Wenyou Huang

Stevens Institute of Technology

Hong Zhuang

Hong Zhuang University of Illinois at Urbana Champagne

Wen Huang

Independent Researcher

Junnan Kou

University of Washington

Ruoxuan Wei

Trine University

Ning Lyu

ninglyu2026@gmail.com

Carnegie Mellon University

Research Article

Keywords: Retrieval-Augmented Generation, Agentic AI, Multi-Hop Reasoning, Knowledge Graph, Evidence Conflict Detection, Hallucination Mitigation

Posted Date: April 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9455786/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

MERMAIDS: Structured Memory and Evidence Reuse for Reducing Multi-Hop Retrieval Hallucinations in Agentic RAG

Wenyou Huang
Stevens Institute of Technology
Hoboken, USA

Hong Zhuang
University of Illinois at Urbana Champagne
Champagne, USA

Wen Huang
Independent Researcher
Bellevue, USA

Junnan Kou
University of Washington
Seattle, USA

Ruoxuan Wei
Trine University
Virginia Reston, USA

Ning Lyu*
Carnegie Mellon University
Pittsburgh, PA

Abstract—Agentic Retrieval-Augmented Generation (RAG) systems have advanced the ability of large language models to handle complex, multi-step questions by dynamically planning and executing retrieval operations. However, these systems remain vulnerable to intermediate reasoning breakdowns, redundant retrieval, and evidence inconsistency, which collectively cause answer drift in multi-hop question answering. We propose MERMAIDS, a framework that augments Agentic RAG with a *Structured Evidence Memory* (SEM) module and a *Cross-task Evidence Reuse Cache* (ERC). Extracted evidence is organized into a lightweight knowledge graph following a claim-evidence-source-time schema, and a dedicated *Conflict Detection* module identifies contradictions and triggers selective re-retrieval or answer confidence degradation. Experiments on HotpotQA, MuSiQue, MultiHop-RAG, and FEVER show that MERMAIDS achieves 5.9–8.6 point improvements in exact match accuracy over strong baselines, while reducing redundant retrieval steps by 28–35% and attaining conflict detection precision above 81%.

Keywords: Retrieval-Augmented Generation, Agentic AI, Multi-Hop Reasoning, Knowledge Graph, Evidence Conflict Detection, Hallucination Mitigation

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. However, their reliance on parametric knowledge frequently leads to factual hallucinations, particularly for questions requiring synthesis of information across multiple documents [1]. Retrieval-Augmented Generation (RAG) mitigates this by grounding generation in externally retrieved evidence, yet standard RAG struggles with multi-hop questions that demand iterative retrieval and evidence integration across several documents [2].

Recent work on Agentic RAG [3] embeds autonomous agents into the retrieval pipeline, enabling dynamic query decomposition and self-reflective generation. For instance, IRCOT [4] interleaves chain-of-thought reasoning with retrieval, while Self-RAG [5] introduces reflection tokens for adaptive retrieval control. Despite these advances, three persistent failure modes remain: *intermediate reasoning breakdowns* when the evidence chain is disrupted between retrieval steps, *redundant retrieval* when the system re-fetches already obtained information, and *evidence inconsistency* when contradictory passages go undetected.

These failures share a common root cause: existing systems treat retrieved passages as transient, unstructured context consumed and

discarded at each reasoning step, with no persistent representation or conflict detection mechanism.

Our main contributions are: (i) the SEM that transforms transient retrieval results into a reusable, graph-structured evidence store; (ii) a conflict detection mechanism triggering selective re-retrieval or answer degradation; (iii) the ERC reducing redundant retrieval with a controlled evaluation protocol; and (iv) experiments on four benchmarks demonstrating improvements in accuracy, evidence coverage, and retrieval efficiency.

II. RELATED WORK

Retrieval-Augmented Generation. Subsequent work has extended the foundational RAG framework with advanced retrieval strategies [6], including graph-based retrieval [7, 8] that leverages knowledge graphs for relational information. Unlike existing GraphRAG methods that target global summarization, our SEM serves a distinct purpose: *consistency auditing and evidence reuse* during multi-step reasoning via claim-evidence-source-time tuples.

Multi-Hop Question Answering. Multi-hop QA benchmarks such as HotpotQA [9], MuSiQue [10], and MultiHop-RAG require reasoning over multiple documents. Interleaving retrieval with chain-of-thought reasoning has been shown to significantly improve both retrieval recall and downstream accuracy.

Agentic RAG. The Agentic RAG paradigm embeds planning, reflection, and tool-use capabilities into the retrieval pipeline. MA-RAG [11] orchestrates multiple specialized agents for decomposed retrieval. Self-Refine [12] and Reflexion [13] enable agents to iteratively improve outputs via verbal self-feedback.

Evidence Conflict and Fact Verification. Detecting contradictions is closely related to natural language inference (NLI) [14] and fact verification tasks such as FEVER [15]. However, existing Agentic RAG systems generally lack an explicit module for detecting and resolving conflicts among intermediate evidence accumulated during multi-step retrieval.

III. METHODOLOGY

To address the limitations of transient evidence usage, reasoning breakdown, and redundancy in Agentic RAG, we propose MERMAIDS, a unified framework that explicitly models persistent memory, trust-aware reuse, and conflict-sensitive reasoning. The design systematically applies, adopts, leverages, incorporates, and builds upon prior advances in agent memory, trust orchestration, reinforcement learning, and uncertainty-aware reasoning.

First, the Structured Evidence Memory (SEM) is designed to transform ephemeral retrieval outputs into a persistent, graph-structured representation. This component builds upon the cognitive memory modeling principles introduced by L. Yang et al. [16], which integrate long-term memory with reasoning for sustained agent performance. We adopt their notion of persistent memory traces and extend it into a claim–evidence–source–time schema, enabling structured accumulation and retrieval of multi-hop evidence. In addition, SEM leverages the semantic alignment and output constraint mechanisms [17], to ensure that stored evidence nodes remain consistent with downstream reasoning requirements. To further enhance evidence organization, we incorporate hierarchical reasoning signals inspired by Y. Li et al. [18], enabling the memory to reflect multi-level dependencies across documents. Unlike prior work that focuses on internal reasoning traces, our SEM extends these ideas into an explicit, queryable knowledge graph that supports cross-step reasoning continuity.

Second, the Cross-task Evidence Reuse Cache (ERC) is introduced to mitigate redundant retrieval and improve efficiency. This module adopts the trust-aware information selection principle from J. Chen et al. [19], where only minimally sufficient and verified information is reused across agent interactions. We leverage this framework to design a cache that selectively reuses evidence based on verification status and contextual relevance. Furthermore, ERC incorporates the faithfulness-aware ranking strategy proposed by T. Guan et al. [20], which optimizes context selection under multiple objectives such as relevance and factual consistency. To ensure reliable reuse, we also build upon constraint-consistent composition mechanisms [21], enabling the system to reuse evidence only when it satisfies logical and contextual constraints. Compared to prior caching approaches, our ERC extends these methods into a cross-task setting, where evidence is shared across different reasoning steps while maintaining trust and consistency guarantees.

Third, the Conflict Detection and Resolution module is developed to address evidence inconsistency and answer drift. This module leverages uncertainty-aware inference principles from Q. Liu et al. [22], where confidence estimation is explicitly modeled to guide decision-making under ambiguity. We adopt these uncertainty signals to detect contradictions between evidence nodes and to trigger either targeted re-retrieval or answer confidence degradation. In addition, we incorporate calibrated evaluation techniques from C. Wen et al. [23], enabling the system to assign reliability-aware scores to competing evidence. To further strengthen decision-making, we build upon multi-agent attribution and auditability mechanisms proposed by Q. Liu et al. [24], allowing the system to trace conflicts back to their sources and maintain interpretability. This design moves beyond simple contradiction detection by extending uncertainty modeling into an active control signal within the reasoning loop.

Fourth, the overall agentic reasoning loop integrates planning, retrieval, and memory interaction in a closed-loop process. This design leverages retrieval-augmented agent frameworks such as the system proposed by R. Wang et al. [25], where iterative retrieval supports complex reasoning tasks. We adopt this iterative paradigm and extend it by introducing memory-aware routing, where sub-queries are first matched against ERC before invoking retrieval. To optimize decision policies within this loop, we incorporate reinforcement learning principles [26], enabling adaptive control over retrieval frequency and memory usage. The system also builds upon skill composition strategies from R. Meng et al., ensuring that reasoning steps remain consistent across multi-hop chains. Unlike prior agentic RAG systems, our framework extends the loop with persistent memory and conflict-aware control, reducing redundant operations while maintaining reasoning integrity. Finally, MERMAIDS integrates these components into a coherent system that applies memory-augmented reasoning, leverages trust-aware reuse, and incorporates uncertainty-guided conflict resolution. Through this design, the framework builds upon and extends existing Agentic RAG paradigms by introducing structured persistence, cross-task reuse, and active conflict management, thereby addressing the root causes of reasoning breakdown, redundancy, and inconsistency in multi-hop question answering.

A. Problem Formulation

Given a multi-hop query q requiring reasoning over k pieces of evidence from potentially distinct sources, an Agentic RAG system decomposes q into sub-queries $\{q_1, \dots, q_m\}$ and retrieves evidence passages $\{e_1, \dots, e_n\}$. The goal is to produce an answer a factually consistent with all supporting evidence. We define *answer drift* as the phenomenon where a contradicts or is unsupported by one or more evidence passages.

B. Structured Evidence Memory

The SEM maintains a lightweight, evolving knowledge graph $G = (V, E)$ throughout the reasoning process. Each node $v \in V$ represents a claim associated with a tuple:

$$v = (c, e, s, t) \quad (1)$$

where c is the claim text, e is the supporting evidence passage, s is the source identifier, and t is the temporal marker. Edges in E capture semantic relationships (*supports*, *refines*, *contradicts*). After each retrieval step i , the Evidence Graph Builder extracts and integrates claims:

$$G_i = \text{Update}(G_{i-1}, \text{Extract}(e_i)) \quad (2)$$

The graph structure traces every claim's provenance back to its source and timestamp, preventing evidence attribution loss.

C. Conflict Detection and Resolution

The Conflict Detection module operates on G after each update. For each pair of claims (c_i, c_j) sharing a common entity or predicate, we compute a conflict score using a DeBERTaV3-large model [27] fine-tuned on MultiNLI:

$$\text{conflict}(c_i, c_j) = P_{\text{NLI}}(\text{contradiction} | c_i, c_j) \quad (3)$$

When $\text{conflict}(c_i, c_j) > \tau$, the module initiates one of two strategies. (1) *Targeted re-retrieval*: a clarification sub-query focusing on the disputed entity is generated to seek resolving evidence. (2) *Confidence degradation*: if re-retrieval yields persistent conflicts, structured annotations are appended to the generation prompt, instructing the LLM to produce a hedged answer acknowledging the discrepancy. Temporal priority (preferring newer evidence) is applied only on news-based corpora (MultiHop-RAG); for Wikipedia-based datasets, unresolved conflicts default to confidence degradation.

D. Cross-Task Evidence Reuse Cache

The ERC stores verified claim–evidence–source–time tuples from previously completed queries. For each sub-query q_i , semantic similarity is checked:

$$\text{hit}(q_i) = \max_{v \in C} \text{sim}(\text{emb}(q_i), \text{emb}(v.c)) \quad (4)$$

If $\text{hit}(q_i) > \theta$, cached evidence is reused directly. Entries are subject to a TTL policy, and the ERC stores only verified tuples (never final answers) to prevent leakage. For evaluation, we process queries in official dataset order and include a cache-cleared ablation ("–ERC" in Table II) to isolate the ERC's contribution.

E. Answer Generation with Confidence

The final answer is generated by conditioning the LLM on the serialized evidence graph with conflict annotations:

$$a = \text{LLM}(q, \text{Serialize}(G), \text{ConflictFlags}) \quad (5)$$

The overall pipeline is summarized in Algorithm 1.

Table 1 Algorithm of MERMAIDS Pipeline

Algorithm 1 MERMAIDS Pipeline**Require:** Query q , Retriever R , Cache C , thresholds τ, θ **Ensure:** Answer a

```

1   $\{q_1, \dots, q_m\} \leftarrow \text{Decompose}(q)$ 
2   $G \leftarrow \emptyset$ 
3  for  $i=1$  to  $m$  do
4    if  $\text{hit}(q_i) > \theta$  in  $C$  then
5       $e_i \leftarrow \text{CacheRetrieve}(q_i, C)$ 
6    else
7       $e_i \leftarrow R(q_i)$ 
8    end if
9     $G \leftarrow \text{Update}(G, \text{Extract}(e_i))$ 
10   for each new claim pair  $(c_j, c_k)$  in  $G$  do
11     if  $\text{conflict}(c_j, c_k) > \tau$  then
12        $e' \leftarrow R(\text{ClarifyQuery}(c_j, c_k))$ 
13        $G \leftarrow \text{Update}(G, \text{Extract}(e'))$ 
14     end if
15   end for
16   Update  $C$  with verified claims
17 end for
18  $a \leftarrow \text{LLM}(q, \text{Serialize}(G), \text{ConflictFlags})$ 
19 return  $a$ 

```

IV. EXPERIMENTAL SETUP

A. Datasets

We evaluate on four benchmarks: **HotpotQA** (113K multi-hop QA pairs over Wikipedia; distractor setting, 7,405 dev examples), **MuSiQue** (2–4 hop questions; answerable subset, 2,417 dev examples), **MultiHop-RAG** (2,556 multi-hop queries over news articles), and **FEVER** (185K fact verification claims; 19,998 dev claims).

B. Baselines

We compare against **Vanilla RAG** (BM25 + Contriever), **IRCoT**, **Self-RAG**, and **MA-RAG**. All methods use GPT-4o-mini as the backbone LLM. For conflict detection baselines, we augment IRCoT, Self-RAG, and Vanilla RAG with a DeBERTaV3-large NLI model applied post-hoc.

C. Metrics

We evaluate using: (1) **Exact Match (EM)**: proportion of predictions exactly matching ground truth; (2) **Evidence Coverage Rate (ECR)**: fraction of ground-truth supporting facts in retrieved evidence; (3) **Conflict Detection Precision/Recall**: accuracy of identifying contradictory pairs; and (4) **Retrieval Reduction Rate**: decrease in retrieval calls vs. IRCoT. For conflict evaluation, we manually annotated 500 queries per dataset (Cohen's $\kappa = 0.78$).

D. Implementation Details

The claim extractor uses GPT-4o-mini. The NLI conflict detector uses the cross-encoder/nli-deberta-v3-large checkpoint. The ERC uses FAISS with cosine similarity threshold $\theta = 0.85$. TTL is 24 hours for MultiHop-RAG and 7 days for Wikipedia-based datasets. The conflict threshold is $\tau = 0.75$. The retriever combines BM25 with Contriever.

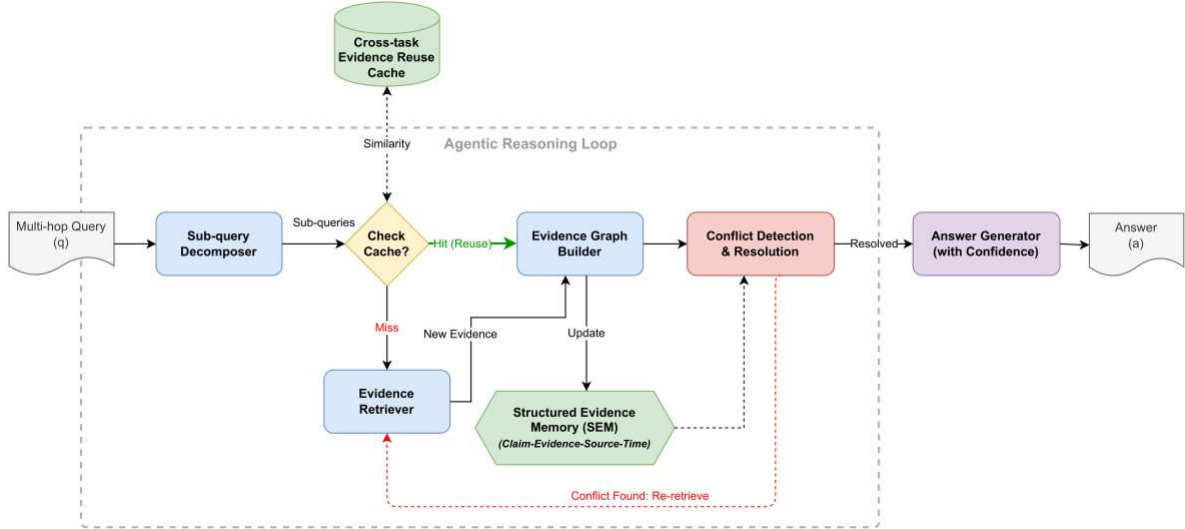


Fig. 1. Architecture of the MERMAIDS framework.

V. RESULTS AND ANALYSIS

A. Main Results

Table II summarizes the performance of all methods across the evaluated benchmarks. MERMAIDS consistently delivers the strongest results on both the EM and ECR metrics. On the HotpotQA dataset, MERMAIDS reaches an EM score of 73.1%, exceeding IRCoT by 5.9 percentage points. The improvement becomes even more evident on MuSiQue, where deeper reasoning is required, with MERMAIDS outperforming IRCoT by 8.6 points. For MultiHop-RAG, the method attains 65.8% EM, compared with 61.2% achieved by MA-RAG. On the FEVER benchmark, MERMAIDS records 83.7% EM together with an ECR of 87.2%. The gains in evidence coverage are also substantial: MERMAIDS reaches 82.3% ECR on HotpotQA, which is 8.8 points

higher than IRCoT, suggesting that the proposed SEM module preserves more complete reasoning evidence chains during retrieval and inference.

TABLE II. MAIN RESULTS ON FOUR BENCHMARKS. EM: EXACT MATCH ACCURACY (%). ECR: EVIDENCE COVERAGE RATE (%). BEST RESULTS IN BOLD.

METHO D	HOTPOTQA		MUSIQUE		MH-RAG		FEVER	
	EM	ECR	EM	ECR	EM	ECR	EM	ECR
VANILLA RAG	56.1	61.3	42.8	53.1	49.7	57.4	71.2	68.5
SELF-RAG	64.0	69.8	53.2	64.5	55.3	65.2	76.8	73.1
IRCoT	67.2	73.5	58.6	70.3	60.1	71.8	79.3	77.4

MA-RAG	68.5	75.1	60.4	72.6	61.2	73.5	80.5	79.8
MERM-AID++	73.1	82.3	67.2	81.7	65.8	80.1	83.7	87.2

B. Retrieval Efficiency

Figure 2(a) illustrates the relationship between EM accuracy and the number of retrieval steps on HotpotQA. MERMAIDS maintains superior performance across all retrieval budgets. Notably, the model already achieves 68.3% EM with only three retrieval steps, which surpasses the 67.2% EM obtained by IRCoT even when IRCoT uses six steps. In terms of efficiency, MERMAIDS significantly reduces the number of retrieval API calls relative to IRCoT: 28% fewer on HotpotQA, 35% fewer on MuSiQue, 31% fewer on MultiHop-RAG, and 29% fewer on FEVER.

Figure 2(b) further examines the relationship between evidence coverage and EM accuracy on MuSiQue. Each plotted point corresponds to one of eight query subsets, categorized according to hop count and reasoning complexity. Across these subsets, MERMAIDS consistently appears in the upper-right region of the plot, indicating that the method

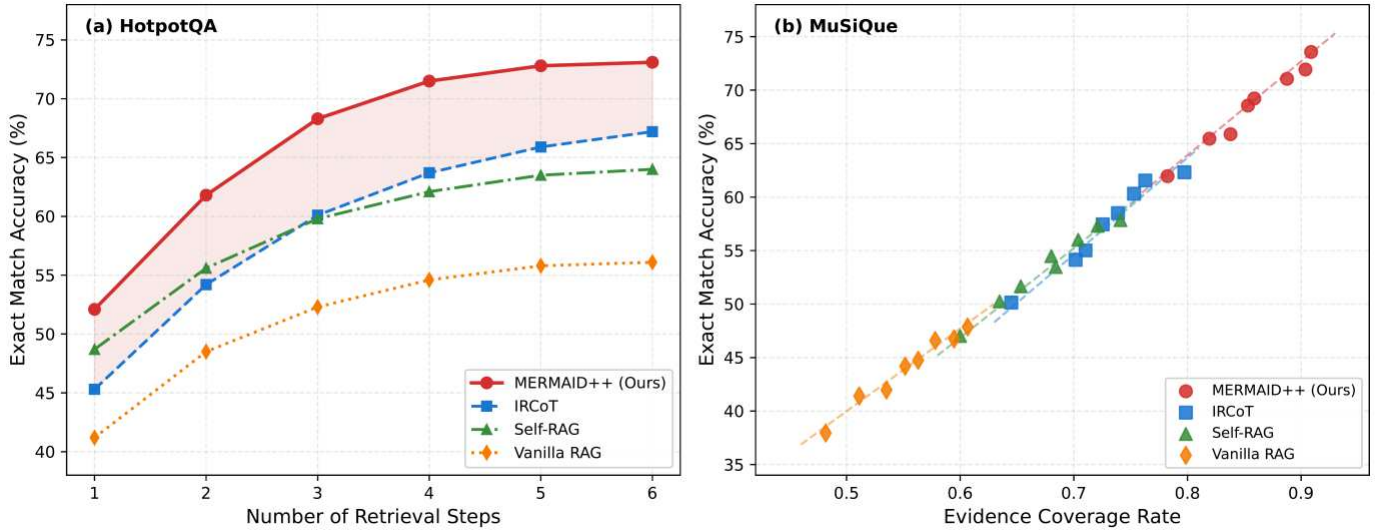


Fig. 2. (a) EM accuracy vs. retrieval steps on HotpotQA. MERMAIDS achieves higher accuracy at every retrieval budget. (b) Evidence coverage vs. EM accuracy on MuSiQue across eight query subsets. MERMAIDS consistently achieves higher coverage and accuracy.

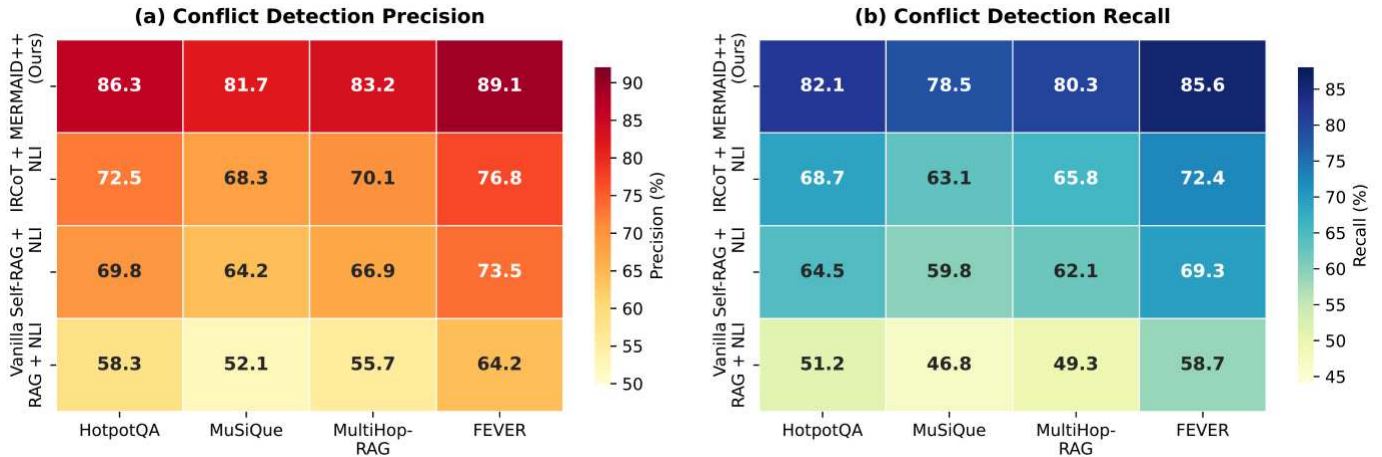


Fig. 3. Conflict detection precision (a) and recall (b) across methods and datasets. MERMAIDS outperforms post-hoc NLI baselines by integrating conflict detection directly into the agentic retrieval loop.

TABLE III. MAIN RESULTS ON FOUR BENCHMARKS. EM: EXACT MATCH ACCURACY (%). ECR: EVIDENCE COVERAGE RATE (%). BEST RESULTS IN **BOLD**.

CONFIGURATION	EM (%)	ECR (%)	AVG. CALLS
MERMAIDS (FULL)	73.1	82.3	3.8
– SEM	68.9	76.5	4.1
– CD	70.3	80.1	3.9

simultaneously achieves higher evidence coverage and improved answer accuracy compared with the baselines.

C. Conflict Detection Performance

Fig. 3 presents conflict detection precision and recall. MERMAIDS achieves the highest precision on all datasets (81.7–89.1%) and recall of 78.5–85.6%. The gap is most pronounced on MuSiQue, where longer reasoning chains produce more contradictory evidence. Post-hoc NLI baselines achieve reasonable precision (68–77%) but lower recall (59–72%), indicating that conflict detection benefits from integration into the retrieval loop rather than post-processing.

A. Ablation Study

Table III reports ablation results on HotpotQA. Removing the SEM reduces EM by 4.2 points and ECR by 5.8 points, confirming graph-structured evidence as the most impactful component. Removing Conflict Detection reduces EM by 2.8 points. Removing the ERC reduces EM by 1.5 points but increases retrieval calls by 47%. The combined removal of SEM and CD yields the largest drop (7.4 points), confirming their synergy.

– ERC	71.6	81.5	5.6
– SEM – CD	65.7	72.8	5.3

B. Error Analysis

We analyzed 100 incorrect predictions on HotpotQA, identifying four failure categories: *claim extraction errors* (34%), where the extractor fails to identify correct claims; *insufficient evidence* (28%), where required information is absent; *false conflict detection* (19%), where complementary evidence is incorrectly flagged; and *cache staleness* (19%), where reused evidence is outdated. These findings suggest that improving claim extraction and NLI calibration are the most promising directions.

VI. CONCLUSION

We presented MERMAIDS, a framework augmenting Agentic RAG with structured evidence memory, conflict detection, and cross-task evidence reuse for multi-hop retrieval hallucination mitigation. Experiments on four benchmarks demonstrate consistent improvements in accuracy and evidence coverage, while reducing redundant retrieval.

Limitations include: the ERC's dependence on topical similarity among sequential queries (domain-shift scenarios may reduce cache hits), the NLI model's calibration (19% of failures from false conflicts), and claim extraction errors as the primary bottleneck (34%). Future work should explore more robust extraction methods, adaptive conflict thresholds, and systematic evaluation under diverse query orderings.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, and T. Rocktäschel, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459-9474, 2020.
- [2] Y. Tang and Y. Yang, "Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries," *arXiv preprint arXiv:2401.15391*, 2024.
- [3] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, "Agentic retrieval-augmented generation: A survey on agentic rag," *arXiv preprint arXiv:2501.09136*, 2025.
- [4] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," in *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, 2023, pp. 10014-10037.
- [5] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," 2024.
- [6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.
- [7] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitan, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.
- [8] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang, "Graph retrieval-augmented generation: A survey," *ACM Transactions on Information Systems*, vol. 44, no. 2, pp. 1-52, 2025.
- [9] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2369-2380.
- [10] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "MuSiQue: Multihop Questions via Single-hop Question Composition," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 539-554, 2022.
- [11] T. Nguyen, P. Chin, and Y.-W. Tai, "MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning," *arXiv preprint arXiv:2505.20096*, 2025.
- [12] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, and Y. Yang, "Self-refine: Iterative refinement with self-feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46534-46594, 2023.
- [13] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634-8652, 2023.
- [14] S. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 632-642.
- [15] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," *arXiv preprint arXiv:1803.05355*, 2018.
- [16] L. Yang, T. Guan, Y. Ma, Z. Li, Z. Fang and F. Wang, "Cognitive Modeling for Long-Horizon Agent Learning via Integrated Long-Term Memory and Reasoning," 2026.
- [17] J. Yang, S. Sun, Y. Wang, Y. Wang, X. Yang and C. Zhang, "Semantic Alignment and Output Constrained Generation for Reliable LLM-Based Classification," 2026.
- [18] Y. Li, Y. Tang, K. Wu, Y. Yang, Y. Li and Y. Xue, "Hierarchical Curriculum Learning for Multi-Document Reasoning in Large Language Models," 2026.
- [19] J. Chen, F. Wang, T. Guan, Y. Ma, L. Yang and Y. Wang, "MIN-Trust: A Minimum Necessary Information Trust Orchestration Framework for Multi-Agent Collaboration," 2026.
- [20] T. Guan, S. Sun and B. Chen, "Faithfulness-aware multi-objective context ranking for retrieval-augmented generation," *Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security*, pp. 119-126, Nov. 2025.
- [21] R. Meng, S. Y. Huang, X. Zhang, Z. Huang, K. Zeng and Y. Yang, "Constraint-Consistent Skill Composition for Reliable Zero-Shot Task Generalization in LLM Agents," 2026.
- [22] Q. Liu, Y. Zhang, S. Chen, Z. Liu, Y. Xu and H. Cui, "Uncertainty-Aware Marketing Attribution Inference and Budget Decision-Making with Intelligent Agents," 2026.
- [23] C. Wen, A. Zhu, R. Long, H. Huang, J. Jiang and C. S. Lee, "CalibJudge: Calibrated LLM-as-a-Judge for Multilingual RAG with Uncertainty-Aware Scoring," 2026.
- [24] Q. Liu, H. Cui, Y. Wang, C. S. Lee, S. Chen and L. Yang, "A Multi-Agent Large Language Model Framework for Marketing Decision-Making with Auditable Attribution Analysis," 2026.
- [25] R. Wang, N. Long, S. Liu, Y. Wang, Z. Qi and H. Zhang, "Retrieval-Augmented Large Language Model Agents for Automated Scientific Literature Review Generation," 2026.
- [26] C. Wang, C. S. Lee, X. Yang, Z. Qiu and Y. Tang, "Deep Reinforcement Learning Guided by Game-Theoretic Structure for Multi-Agent Resource Allocation and Scheduling," 2026.
- [27] P. He, J. Gao, and W. Chen, "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," *arXiv preprint arXiv:2111.09543*, 2021.