TRUST-RAG: AGENTIC RETRIEVAL-AUGMENTED GENERATION WITH EVIDENCE VERIFICATION AND CONSISTENCY-AWARE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) improves large language models by incorporating relevant external information during answer generation. However, retrieved evidence may be irrelevant, unreliable, or contradictory, potentially leading to incorrect reasoning and unsupported responses. We propose Trust-RAG, an agentic RAG framework designed to improve evidence reliability and consistency in multi-hop question answering. Trust-RAG comprises three specialized agents, Retrieval, Verification, and Consistency coordinated by an LLM-based controller. The Retrieval Agent adaptively selects the retrieval query and retrieval depth, the Verification Agent evaluates and filters candidate documents, and the Consistency Agent identifies contradiction-free evidence and determines whether additional retrieval is required. The controller coordinates these decisions and initiates query reformulation and re-retrieval when the available evidence is insufficient. Trust-RAG combines BM25 and dense retrieval within a hybrid retrieval system, verifies evidence using relevance and natural language inference (NLI) signals, and applies consistency-graph reasoning before answer generation. Qwen2.5-7B-Instruct is used for agent decision-making and final answer generation. We evaluate Trust-RAG on HotpotQA and GSM8K. On the full HotpotQA distractor validation set ($n = 7{,}405$), Trust-RAG achieves an Exact Match score of 0.423 and an F1 score of 0.535, while the selected consistent evidence retains all gold supporting documents in 58.7% of examples. On the full GSM8K test set ($n = 1{,}319$), Trust-RAG achieves a numerical Exact Match score of 0.830. These results demonstrate the applicability of Trust-RAG to both evidence-intensive multi-hop question answering and mathematical reasoning.

1 INTRODUCTION

Large language models (LLMs) can answer questions using knowledge acquired during training, but their responses may be outdated, unsupported, or inconsistent with available evidence Lewis et al. (2020). Retrieval-augmented generation (RAG) addresses this limitation by incorporating external documents at inference time Lewis et al. (2020). However, retrieval alone does not guarantee reliable generation. Retrieved passages may be irrelevant, incomplete, contradictory, or only superficially related to the query Wang et al. (2025). These challenges are particularly significant in multi-hop question answering, where correct responses require integrating evidence distributed across multiple documents Yang et al. (2018). Providing all retrieved content directly to an LLM can introduce distracting or conflicting information, leaving the model responsible for determining which evidence should be trusted Shi et al. (2023).

Many RAG systems follow a retrieve-then-generate paradigm in which evidence is collected before answer generation. FLARE dynamically retrieves additional documents during generation when predicted content is uncertain Jiang et al. (2023), while IRCoT interleaves retrieval with chain-of-thought reasoning for multi-step questions Trivedi et al. (2023). Adaptive-RAG selects among no-retrieval, single-step, and iterative retrieval strategies according to question complexity Jeong et al. (2024). Self-RAG uses reflection tokens to determine when retrieval is needed and when retrieved evidence or generated responses should be critiqued Asai et al. (2024). These approaches improve retrieval flexibility, but their primary control signals are typically uncertainty, question complexity,

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.

or self-reflection. Trust-RAG differs by enabling the Retrieval Agent to adapt retrieval depth across both questions and retrieval rounds.

Post-retrieval approaches have focused on mitigating the effects of irrelevant or unreliable evidence. RECOMP compresses retrieved documents and can omit augmentation when the retrieved evidence is considered uninformative Xu et al. (2024). RAGTruth demonstrates that RAG systems can still produce claims that are unsupported by or contradictory to retrieved evidence Niu et al. (2024). Astute RAG addresses imperfect retrieval and knowledge conflicts through reliability-aware consolidation Wang et al. (2025), whereas MAIN-RAG employs multiple agents and inter-agent consensus to filter noisy documents Chang et al. (2025). Trust-RAG builds on this direction by explicitly separating document-level verification from claim-level consistency analysis. It further removes lower-confidence contradictory claims before constructing the final evidence set used for reasoning and generation.

Other studies have explored structured evidence organization and source traceability. ALCE evaluates whether generated statements are adequately supported by their cited sources Gao et al. (2023). HippoRAG combines knowledge graphs with graph-based retrieval to support multi-hop question answering Gutiérrez et al. (2024). In contrast, Trust-RAG constructs a question-specific graph from verified documents, preserves the provenance of extracted claims, and explicitly represents support, contradiction, heuristic causal, and shared-entity relationships. The resulting graph is used to identify consistent evidence and to determine whether additional retrieval is necessary.

We propose Trust-RAG, an agentic retrieval-augmented generation framework composed of three specialized agents coordinated by an LLM-based controller. The Retrieval Agent combines BM25 and dense retrieval through weighted ranking and adaptively selects a retrieval depth from $k \in \{5, 7, 9\}$. The Verification Agent scores candidate documents using relevance, reliability, and contradiction signals and filters weak evidence before it reaches the reasoning stage. The Consistency Agent then extracts claims and entities and constructs an evidence consistency graph containing entity associations as well as support, contradiction, and heuristic causal relationships identified through lexical cues. This graph is used to select a contradiction-free reasoning path together with a broader set of consistent evidence for answer generation. When the available evidence is insufficient or contains unresolved conflicts, the controller can refine the query and initiate an additional retrieval round. The final reader therefore receives verified and consistency-checked evidence rather than an unfiltered retrieval set.

This work makes three main contributions. First, we introduce an adaptive multi-agent RAG architecture that separates retrieval, verification, and consistency decisions. Second, we develop an evidence consistency graph for contradiction handling, multi-document reasoning, and source traceability. Third, we evaluate Trust-RAG on HotpotQA Yang et al. (2018) and GSM8K OpenAI (2021), demonstrating its applicability to both multi-hop question answering and mathematical reasoning.

2 Proposed Trust-RAG Framework

2.1 Framework Overview

Trust-RAG processes each question using three specialized agents coordinated by an LLM-based controller. The controller observes the current state of the system and determines which action should occur next. The Retrieval Agent selects the query and retrieval depth, the Verification Agent determines which retrieved documents should be retained, and the Consistency Agent evaluates the evidence using the consistency graph and determines whether it is sufficient. If additional evidence is required, the controller can initiate query refinement and another retrieval round. Up to two additional retrieval rounds are allowed before the best available answer is returned.

Figure 1 illustrates the end-to-end workflow of Trust-RAG. Given an input question, the controller first invokes the Retrieval Agent, which performs hybrid BM25–dense retrieval and adaptively selects the retrieval depth based on agreement between the two retrievers. The retrieved candidates are then passed to the Verification Agent, which evaluates their relevance, reliability, and contradiction signals and filters low-confidence evidence. The retained documents are subsequently processed by the Consistency Agent, which extracts claims and entities, constructs the consistency graph, removes contradictory reasoning paths, and selects the strongest consistent evidence path. Based on

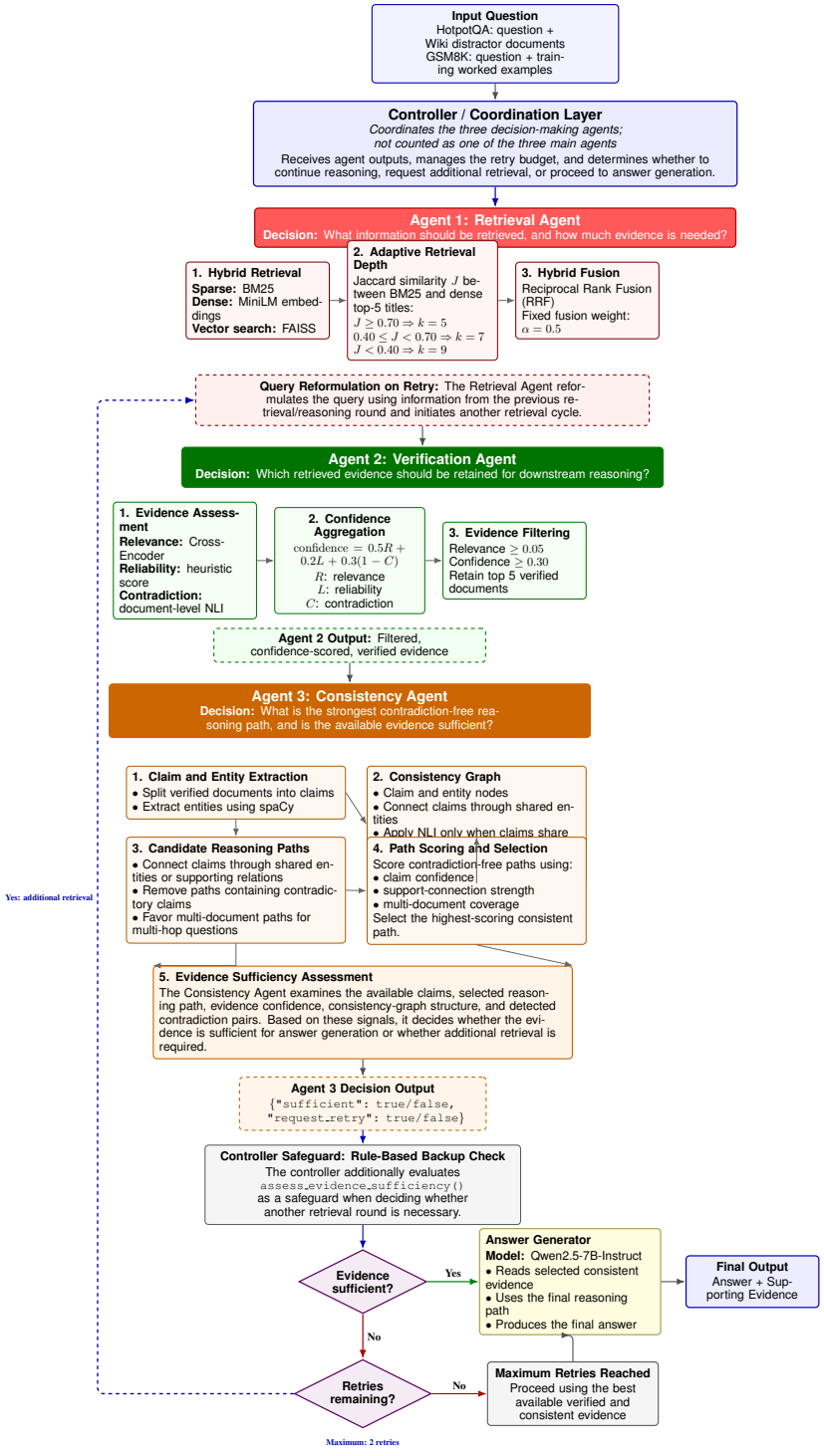

Figure 1: Overview of the proposed Agentic Trust-RAG framework. The Retrieval Agent performs hybrid sparse–dense retrieval and adaptively selects the retrieval depth; the Verification Agent scores and filters candidate evidence using relevance, reliability, and contradiction signals; and the Consistency Agent constructs a claim-level consistency graph, removes contradictory reasoning paths, and evaluates evidence sufficiency. The controller coordinates the three agents and triggers query reformulation and re-retrieval when necessary. Once sufficient consistent evidence is obtained, Qwen2.5-7B-Instruct generates the final answer from the selected evidence.

the resulting evidence confidence and graph structure, the Consistency Agent determines whether the available evidence is sufficient. If additional evidence is required and the retry budget remains, the controller triggers query reformulation and another retrieval round; otherwise, the selected evidence is passed to Qwen2.5-7B-Instruct for final answer generation. This iterative design prevents raw retrieved evidence from being directly exposed to the answer generator and allows unreliable or conflicting information to be filtered before generation.

2.2 Query Analysis and Reformulation

Trust-RAG performs query reformulation when the available evidence is insufficient or conflicting. The first retrieval round typically uses the original question. After verification and consistency analysis, the Consistency Agent determines whether additional evidence is required. If so, the controller initiates query refinement.

The reformulation process considers the original question together with detected contradiction pairs and shared entities identified in the consistency graph. Documents associated with detected conflicts can also be excluded from the next retrieval round. Using this information, the Retrieval Agent refines the query and selects a new retrieval depth $k \in \{5, 7, 9\}$ to target missing or conflicting evidence. Up to two additional retrieval rounds can be performed when necessary.

2.3 Adaptive Retrieval Agent

The Retrieval Agent determines both the retrieval query and the amount of evidence to retrieve, selecting a retrieval depth from $k \in \{5, 7, 9\}$. Retrieval is performed using a hybrid combination of BM25 and dense retrieval with `all-MiniLM-L6-v2` embeddings indexed using FAISS. The rankings are combined using weighted Reciprocal Rank Fusion (RRF) with $\alpha = 0.5$.

The Retrieval Agent selects a retrieval depth from $k \in \{5, 7, 9\}$ based on the question, retrieval round, and the available contradiction signals. BM25 dense overlap is additionally computed as a retrieval agreement signal. The Jaccard similarity is calculated on their top-five retrieved titles, with higher agreement favoring a smaller retrieval depth and lower agreement favoring a larger depth. Candidate documents are passed to the Verification Agent.

2.4 Evidence Verification Agent

The Verification Agent evaluates the retrieved documents before they are used for reasoning. Candidate documents are assessed using relevance, reliability, and contradiction signals. These signals are combined into an evidence-confidence score:

$$\text{confidence} = 0.5R + 0.2L + 0.3(1 - C)$$

Default filtering retains up to five documents with relevance scores of at least 0.05 and confidence scores of at least 0.30. The Verification Agent then decides which document titles to retain, optionally overriding the default filtering decision. The retained documents form the verified evidence passed to the consistency graph.

2.5 Evidence Consistency Graph Construction

Verified documents are divided into individual claims, and the named entities are extracted using spaCy. A consistency graph is then constructed from claim ad entity nodes. Claims are connected through shared entities, while NLI is used to identify support and contradiction relationships between claims that share an entity. A threshold of 0.60 is used for support and contradiction relations. The resulting graph provides a structured representation of relationships and conflicts within the verified evidence.

2.6 Consistency-Aware Graph Reasoning

The consistency graph is used to identify a contradiction-free reasoning path and a broader set of consistent evidence. Candidate reasoning paths are evaluated using claim confidence, supporting relationships, and multi-document coverage. Paths containing contradictory claims are excluded.

The Consistency Agent then examines the resulting claims, reasoning path, evidence confidence, and detected contradiction pairs to determine whether the available evidence is sufficient. It can also identify which claims should be trusted and whether another retrieval round should be requested.

2.7 Evidence-Grounded Answer Generation

Answer generation is the final reading stage of Trust-RAG, where the question and selected consistent evidence are used to produce an evidence-based answer. This stage is reached when the controller determines that sufficient verified and consistency-checked evidence is available. If the evidence remains insufficient and additional retrieval rounds are available, the controller initiates query refinement and re-retrieval.

At this stage, **Qwen2.5-7B-Instruct** functions as the reader and answer generator. Once the controller determines that the system is ready to answer, the reader receives the selected consistent evidence rather than the raw retrieved documents. It reads this evidence together with the question and generates the final answer. The reader does not select documents, resolve conflicts, or determine whether additional retrieval is required. These decisions are handled by the controller and the three specialized agents.

Trust-RAG does not use strict abstention by default. Confidence from the verification and consistency stages is used to determine the reliability of the available evidence. If the maximum retrieval budget is reached, the system returns the best available answer from the strongest consistent evidence rather than refusing to answer. Strict abstention could be incorporated in future work for applications where answering under uncertainty is undesirable.

3 Experimental Methodology

3.1 Dataset

Trust-RAG is evaluated on HotpotQA and GSM8K. HotpotQA serves as the primary multi-hop question-answering dataset. Experiments are conducted on questions accompanied by approximately ten Wikipedia paragraphs, including two gold supporting pages and several distractors. A randomly selected subset of 2,000 training examples, generated using seed 42, is used to tune the hyperparameters. The final evaluation is conducted on the validation set that contains 7,405 examples. GSM8K is used to examine mathematical reasoning. This dataset does not provide retrieval passages; therefore, the training questions and their worked solutions are converted into a retrieval corpus. Final evaluation is conducted on the full GSM8K test set containing 1,319 examples.

3.2 Baseline

Retrieval performance is compared across BM25, dense retrieval, and fixed hybrid retrieval. These methods are evaluated on the full HotpotQA validation set using Title Recall@7. The complete Trust-RAG framework is evaluated separately for end-to-end answer quality on HotpotQA and GSM8K.

3.3 Implementation details

The proposed Trust-RAG framework contains three main agents; a Retrieval Agent, a Verification Agent, and a Consistency Agent. These agents use the retrieval, verification, and consistency mechanisms as tools and make decisions within their respective stages. A controller coordinated the three agents and determines the next action in the workflow, including whether additional retrieval is required.

Dense retrieval is implemented using `all-MiniLM-L6-v2` embeddings indexed with FAISS. Evidence verification uses `ms-marco-MiniLM-L-6-v2` for relevance scoring and `nli-MiniLM2-L6-H768` for contradiction detection. The named entities are extracted using the `spaCy model en_core_web_sm`. `Qwen2.5-7B-Instruct` is used for answer generation and for the decision-making steps of the agents and controller.

The hybrid retriever used a fixed fusion weight of $\alpha = 0.5$ and a base retrieval depth of $k = 7$. During inference, the Retrieval Agent can select $k \in \{5, 7, 9\}$ and reformulate the retrieval query when additional evidence is needed. The Verification Agent evaluates the retrieved documents and keeps up to five documents with relevance scores of t least 0.05 and confidence scores of at least 0.30, where

$$\text{confidence} = 0.5 \cdot \text{relevance} + 0.2 \cdot \text{reliability} + 0.3 \cdot (1 - \text{contradiction}).$$

The Consistency Agent uses the consistency graph to evaluate the available claims and determine whether the evidence is sufficient. Claim-level support and contradiction relationships use an NLI threshold of 0.60. If the evidence is insufficient, the controller can initiate another retrieval round. Up to two additional retrieval rounds are allowed before the best available answer is returned.

3.4 Evaluation metrics

ForHotpotQA, answer quality is evaluated using Exact Match (EM) and token-level F1 following standard answer normalization. Retrieval performance is measured using Title Recall@7. For the complete Trust-RAG pipeline, additional metrics include gold-title coverage within the selected evidence and reasoning path, together with the timeout rate.

For GSM8K, performance is measured using numerical Exact Match, which compares the extracted predicted number with the gold answer. Token-level F1 is not applied because GSM8K answers are evaluated as single numerical values.

4 Results and Analysis

4.1 Retrieval performance

Trust-RAG retrieves evidence with a hybrid of BM25 and dense retrieval with a variable retrieval depth. The fusion weight and base depth are selected on a 2,000 example HotpotQA training development subset. During inference, the Retrieval Agent and controller select a depth from $k \in \{5, 7, 9\}$ for each question. BM25 and dense rankings are fused under that chosen depth.

Retrieval performance on HotpotQA is evaluated using Title Recall@7. BM25, dense retrieval, and fixed hybrid retrieval are compared to measure the contribution of hybrid fusion.

Table 1: HotpotQA fixed retrieval performance on the full validation set.

Retriever	Title Recall@7
BM25	0.869
Dense	0.898
Fixed hybrid	**0.923**

GSM8K does not provide gold retrieval passages, so retrieval recall cannot be measured directly. Its retrieval contribution is instead evaluated through numerical answer accuracy and is reported in the answer-quality results.

4.2 Answer Quality

Answer quality is evaluated using Exact Match (EM), token-level F1 on HotpotQA, and numerical Exact Match on GSM8K. Trust-RAG is evaluated on the full HotpotQA distractor validation set and the full GSM8K test set.

4.3 Unsupported Generation and Evidence Reliability Analysis

HotpotQA and GSM8K do not directly identify hallucination in the generated answers. Therefore, data set-specific proxy measures are used to evaluate the reliability of evidence and the generation of unsupported data. In HotpotQA, gold supporting evidence that remains after filtering the distractor

Table 2: Answer quality of Trust-RAG on HotpotQA and GSM8K.

Dataset	EM	F1
HotpotQA	0.423	0.535
GSM8K	0.830	–

documents is measured, as reported in the following subsection. In GSM8K, incorrect predictions are classified as *copy errors* when they match the final answer of a retrieved example and as *novel errors* otherwise. The corresponding rates are *0.8%* and *15.3%*, respectively. Figure 2 provides an overview of Trust-RAG performance on GSM8K, showing strong numerical answer accuracy alongside relatively low copy-error and timeout rates.

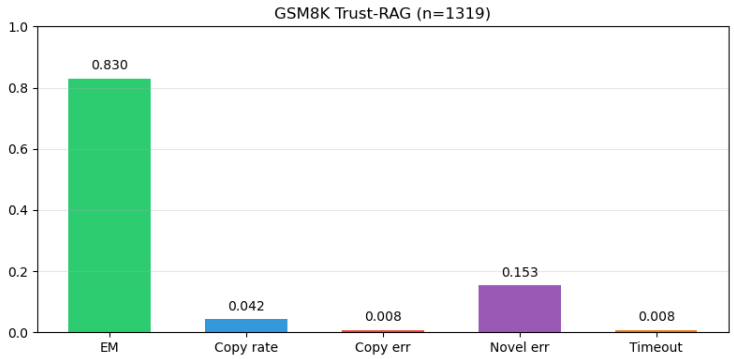

Figure 2: End-to-end Trust-RAG results on the full GSM8K test set ($n = 1{,}319$).

4.4 Evidence Reliability and Coverage

In HotpotQA, evidence preservation is evaluated by whether all gold supporting titles remain after evidence processing. The contradiction-free evidence set retained the complete set of gold titles in 58.7% of the examples, while the selected reasoning path did so in 22.2%. These results illustrate the trade off between preserving complete multi document evidence and restricting the reader to a narrower set of consistent claims. Figure 3 illustrates the trade-off between answer quality and evidence preservation on HotpotQA, comparing EM and F1 with gold-title coverage in the selected reasoning path and the broader evidence set.

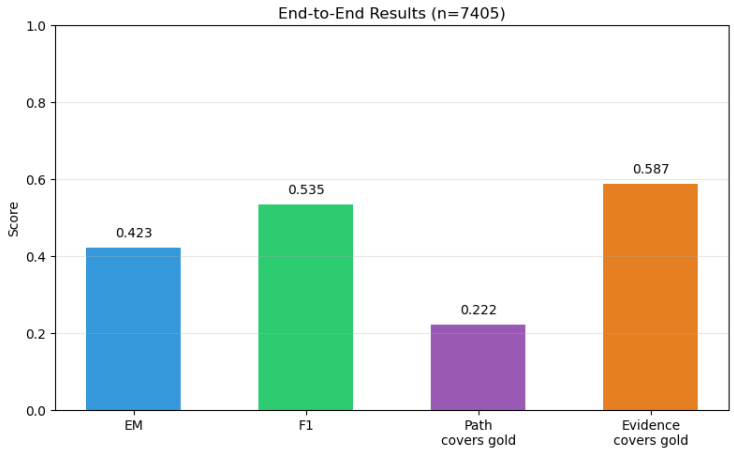

Figure 3: End-to-end Trust-RAG results on the full HotpotQA distractor validation set ($n = 7{,}405$).

4.5 CONTRADICTION-HANDLING RESULTS

Trust-RAG is able to identify contradictions at the claim level. The lower-confidence claim for each conflicting pair is removed before the final selection of evidence. Detected contradiction pairs can trigger re-retrieval and query reformulation. Contradiction handling is evaluated within the complete piepline. Therefore, the reported answer quality and evidence coverage results reflect its use but do not isolate its individual contribution.

4.6 MULTI-HOP REASONING RESULTS

Multi-hop reasoning is evaluated on HotpotQA by examining the evidence path constructed across documents. The selected path spans at least two distinct document titles in 80.2% of the validation and contains all gold supporting titles in 22.2%. The selected reasoning path frequently spans multiple documents, although its complete gold-title coverage remains limited.

4.7 EFFICIENCY AND COMPUTATIONAL OVERHEAD

Evidence verification, consistency-graph construction, and conditional re-retrieval introduce additional computation. Retrieval depth is bounded to $(k \in \{5, 7, 9\}$. Under per-example limits of 420 seconds for HotpotQA and 180 seconds for GSM8K, the timeout rates were 0.1% and 0.8%, respectively. The complete pipeline was able to successfully process more than 99% of both evaluation sets.

5 CONCLUSION AND FUTURE DIRECTION

Trust-RAG is a coordinated three-agent framework that integrates adaptive hybrid retrieval, evidence verification, and claim-level consistency analysis. Its consistency graph organizes retrieved claims and filters contradictory evidence before answer generation. On HotpotQA, fixed hybrid retrieval achieved a Title Recall@7 of 0.923, while the complete framework achieved an Exact Match score of 0.423 and an F1 score of 0.535. The contradiction-free evidence set retained all gold supporting titles in 58.7% of the validation examples. On GSM8K, Trust-RAG achieved an Exact Match score of 0.830. These results demonstrate the framework's ability to combine adaptive retrieval with structured evidence processing across both multi-hop question answering and mathematical reasoning.

Future work will evaluate Trust-RAG on additional question-answering, reasoning, and domain-specific datasets to assess its generalizability beyond HotpotQA and GSM8K. The retrieval-depth policy could also be extended beyond the fixed choices of $k \in \{5, 7, 9\}$, which were designed for the approximately ten candidate documents provided in HotpotQA. A learned policy could dynamically select k based on factors such as the number of available documents, question complexity, agreement between retrievers, and the expected value of additional evidence. Similarly, filtering and contradiction thresholds could be made adaptive rather than fixed. Another direction is to enable Trust-RAG to detect ambiguous questions and request clarification before initiating retrieval. Additional future work includes improving reasoning-path coverage, isolating the contributions of contradiction handling and re-retrieval through ablation studies, incorporating abstention when sufficient consistent evidence cannot be identified, and reducing the computational cost of evidence verification and consistency-graph construction.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, grammatical corrections, and improving the clarity and readability of the manuscript. Generative AI tools were not used for developing the methodology, conducting experiments, analyzing data, generating results, or drawing scientific conclusions.

All AI-assisted edits were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2607–2622. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.131. URL https://aclanthology.org/2025.acl-long.131/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong 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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

OpenAI. GSM8K: Grade school math 8k dataset. https://github.com/openai/grade-school-math, 2021. Official dataset repository.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish 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)*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bda88ed2892f5e61c9a9bf215c566913-Paper-Conference.pdf.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher 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*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.