Answer with Evidence: Dual-Path Retrieval-Augmented Generation with Evidence-Grounded Debate

Anonymous ACL submission

Abstract

While Retrieval-Augmented Generation (RAG) can largely reduce large language model hallucinations, ensuring consistent output reliability remains a critical challenge due to retrieval inaccuracies and prior knowledge bias. Conventional RAG architectures often rely on single-pass retrieval mechanisms that lack internal verification, leaving them unable to detect or rectify unsupported claims. To address this, we introduce Answer with Evidence (AwE), a prompt-only, dual-path framework that enforces rigorous grounding through adversarial verification. AwE separates reasoning into two distinct paths: a Knowledge-First path, which elicits an initial answer from the model's parametric memory, retrieves supporting or contradicting passages, and then re-generates a refined answer; and a Retrieval-First path, which performs two iterative retrieval hops and produces an independent answer. Each path produces a candidate answer along with the passages it references. The same frozen LLM conducts a single-round, evidence-grounded debate where each candidate is restricted to citing spans from its own passage set. A referee model then selects the better-supported answer, ensuring final outputs are strictly grounded in retrieved evidence. Experiments on five knowledge-intensive benchmarks showed consistent gains in exact-match accuracy and substantially fewer unattributed statements, confirming that citation-constrained self-debate can be injected at inference time. Our code will be open-sourced upon publication.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse tasks, yet they frequently suffer from hallucinations—generating plausible but factually incorrect content (Ji et al., 2023; Huang et al., 2025; Achiam et al., 2023; Touvron et al., 2023). This limitation becomes particularly problematic in knowledge-intensive scenarios where factual accuracy is paramount. Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to mitigate hallucinations by incorporating external knowledge retrieval into the generation process (Gao et al., 2023; Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021). However, RAG systems introduce a critical vulnerability: the quality of generated responses heavily depends on the accuracy and relevance of retrieved documents. When retrieval fails to obtain correct or sufficient information, or when generation deviates from retrieved evidence, the system can produce what we term "compounded hallucinations"—errors that are more misleading than the original model hallucinations(Hu et al., 2025).

Recent efforts to address RAG limitations have focused on either improving retrieval quality (Trivedi et al., 2023; Shao et al., 2023) or enhancing generation robustness (Asai et al., 2023; Kim et al., 2024). Iterative retrieval methods (Jiang et al., 2023; Yu et al., 2024) attempt to refine search queries through multi-round interactions, while reflection-based approaches (Asai et al., 2023) incorporate self-evaluation mechanisms during generation. However, these methods exhibit two fundamental weaknesses: (1) They treat retrieval and generation as sequential stages without ensuring coherent integration, and (2) They lack mechanisms to verify that generated content strictly adheres to retrieved evidence. Consequently, models may still produce hallucinations by either over-relying on potentially flawed parametric knowledge or inadequately utilizing retrieved information.

We observe that reliable question answering requires a delicate balance between leveraging model's internal knowledge and strictly adhering to external evidence. Model parametric knowledge, while potentially outdated or biased, provides valuable reasoning patterns and contextual understanding. Meanwhile, retrieved documents

offer factual grounding but may be incomplete or contain irrelevant information. This observation motivates our key insight: effective hallucination mitigation requires both comprehensive evidence gathering and rigorous evidence adherence, supported by mechanisms that can adjudicate between competing knowledge sources while maintaining factual accountability.

To this end, we propose Answer with Evidence (AwE), a novel dual-path framework that enforces strict evidence-grounded generation through an innovative debate mechanism. Our approach comprises two complementary retrieval-generation paths designed to maximize evidence coverage while maintaining generation fidelity. The **Knowledge-First Path** begins by eliciting the model's parametric knowledge to produce an initial answer, which then guides targeted retrieval of supporting or contradicting evidence. This path effectively leverages model's internal understanding while remaining open to external verification. The **Retrieval-First Path** performs iterative retrieval, where initial retrieved information informs subsequent retrieval rounds to gather comprehensive contextual evidence. This path ensures thorough exploration of external knowledge sources. Both paths generate candidate answers under strict constraints that all claims must be substantiated by retrieved documents, promoting answer accountability.

The critical innovation of AwE lies in its evidence-grounded debate phase, where candidate answers from both paths undergo rigorous comparison. Unlike conventional debate mechanisms that allow unrestricted argumentation (Du et al., 2024; Liang et al., 2024), our debate enforces an *evidence-constraint* principle: all arguments and counter-arguments must cite specific passages from their respective retrieved document sets. This design ensures that the debate process remains grounded in verifiable evidence while preventing the introduction of unsubstantiated claims. The debate proceeds through structured rounds where each path attempts to demonstrate the superiority of its answer based on evidence quantity, quality, and consistency. A judge mechanism evaluates the evidential support for each position and selects the most reliable answer.

We conduct comprehensive experiments on six knowledge-intensive benchmarks spanning open-domain question answering, multi-hop reasoning, and commonsense reasoning tasks. Results demonstrate that AwE achieves substantial improvements over strong RAG baselines, with particularly notable gains on complex reasoning tasks requiring multi-step inference. Our analysis reveals that the dual-path design effectively captures complementary aspects of knowledge, while the evidence-grounded debate significantly reduces unsubstantiated claims. Furthermore, AwE maintains strong performance across varying retrieval qualities, indicating robustness to imperfect retrieval conditions.

Our main contributions are as follows:

- We propose AwE, a novel dual-path framework that balances model knowledge with strict evidence adherence to mitigate compounded hallucinations in RAG systems.
- We introduce an evidence-grounded debate mechanism that enforces factual accountability by requiring all arguments to be substantiated by retrieved documents.
- We demonstrate through extensive experiments that AwE significantly improves factual accuracy while maintaining strong reasoning capabilities across diverse knowledge-intensive tasks.

2 Related Work

Retrieval-augmented generation. Early retrieve-then-read systems (Lewis et al., 2020; Izacard and Grave, 2021) obtain a single document set and generate once. To answer multi-hop questions, iterative RAG frameworks interleave retrieval with reasoning steps (Trivedi et al., 2023), refine queries with previously generated content (Jiang et al., 2023), or learn when to stop (Yu et al., 2024). Despite improved recall, none of these methods hard-constrain the final answer to be entailed by the retrieved passages; models routinely insert parametric details (Ji et al., 2023). AwE closes this accountability gap by forcing each path to produce a closed-bundle (answer + evidence) before any comparison happens.

Hallucination mitigation. Post-hoc attribution systems (Kim et al., 2024) generate freely and afterwards search for supporting spans, often suffering confirmation bias. Training-time solutions such as Self-RAG (Asai et al., 2023) or RLHF (Ouyang et al., 2022) introduce new parameters or curated preference data, hindering zero-shot domain

transfer. Inference-time approaches like chain-of-verification (Dhuliawala et al., 2024) still rely on parametric knowledge for the verification step. In contrast, AwE imposes prospective citation constraints—every claim must quote its evidence before it reaches the judge—without any additional training.

Faithful and controllable QA. The requirement that model outputs be *entailed* by provided context has been studied for reading comprehension (Jia and Liang, 2017) and extractive QA (Durmus et al., 2020). In the open-domain setting, attributed QA (Bohnet et al., 2022) asks models to append citations, yet still allows unconstrained drafting. Controllable generation work adds hard decoding constraints (Krishna et al., 2023; Dziri et al., 2021) but requires specialised architectures. We instead achieve faithfulness through prompt-level restrictions inside a vanilla LLM, keeping the solution training-free and plug-and-play.

Multi-agent debate. Unconstrained debate improves reasoning (Du et al., 2024; Liang et al., 2024) but can amplify parametric errors (Zheng et al., 2024). Recent "evidence-based" games (Lin et al., 2023; Hu et al., 2025) relax evidentiary granularity or permit single-shot justification. AwE enforces strict span-level citation: all arguments—support or rebuttal—must reference concrete sentences from the path's own retrieved set, preventing the debate itself from introducing new hallucinations.

Research gap. Prior work treats retrieval, grounding, and debate as independent layers. AwE integrates (i) complementary dual-path evidence collection, (ii) prospective citation constraints, and (iii) span-level debate adjudication into one training-free pipeline, yielding consistently more faithful answers across diverse open-domain tasks.

3 Methodology

In this section, we present Answer with Evidence (AwE), a dual-path framework that enforces strict evidence adherence through an evidence-grounded debate mechanism. As illustrated in Figure 1, AwE consists of three main components: (1) Knowledge-First Path that leverages model's parametric knowledge while maintaining evidence accountability, (2) Retrieval-First Path that performs comprehensive evidence gathering through iterative retrieval, and (3) Evidence-Grounded Debate that adjudicates between candidate answers under strict evidentiary constraints.

3.1 Problem Formulation

Given a question q, our goal is to generate an answer a that is strictly grounded in retrieved evidence while maximizing factual accuracy. Unlike traditional RAG systems that generate answers based on a single set of retrieved documents $D = \{d_1, d_2, ..., d_k\}$, our approach, AwE, employs two complementary retrieval paths to obtain two distinct document sets, D_1 and D_2. Each set supports a candidate answer generated under strict evidence constraints. The final answer is selected through an evidence-grounded debate, where both candidate answers are evaluated against the retrieved documents, ensuring that all claims are substantiated by specific document citations.

3.2 Knowledge-First Path

The Knowledge-First Path leverages the model's parametric knowledge as a starting point while ensuring ultimate evidence accountability. This path consists of three sequential stages:

Stage 1: Knowledge Elicitation We first prompt the model to generate an initial answer based solely on its parametric knowledge:

$$a_{init} = \mathcal{M}(q) \tag{1}$$

where $\mathcal{M}$ represents the language model. This stage captures the model's inherent understanding of the query, which may contain accurate information, hallucinations, or partial knowledge.

Stage 2: Targeted RAG Using the initial answer a_{init}, we construct a retrieval query that combines the original question with the generated knowledge:

$$q_{kf} = \textsc{Concat}(q, a_{init}) \tag{2}$$

This composite query guides retrieval toward documents that either support or contradict the model's initial response. We retrieve the top-k documents based on relevance scoring:

$$D_{init} = \textsc{Retrieve}(q_{kf}, k) \tag{3}$$

With retrieved documents D_{init}, we generate the final answer for this path:

$$A_k = \mathcal{M}(q, D_{init}) \tag{4}$$

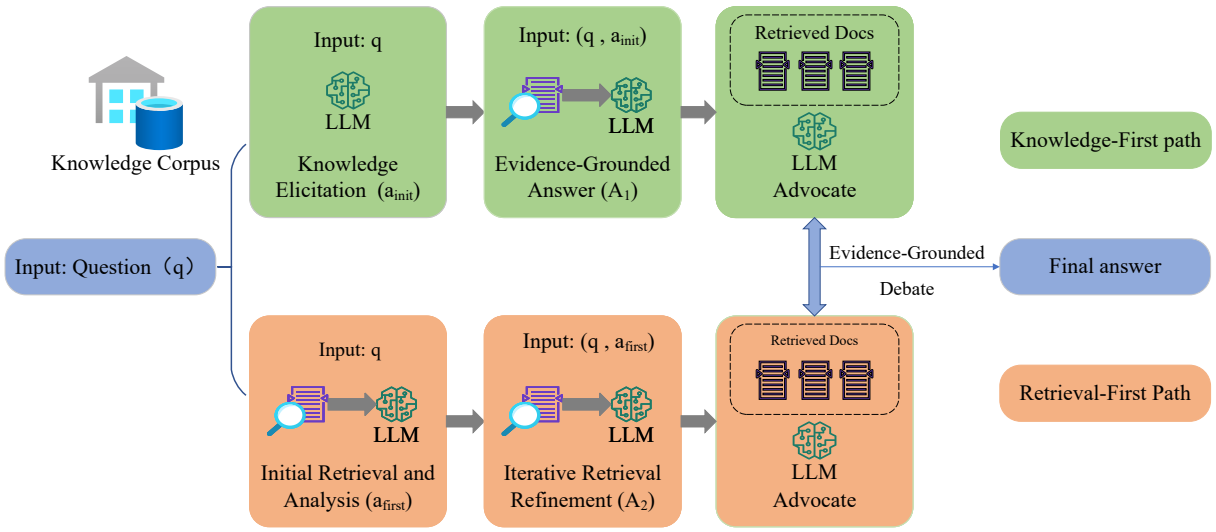

Figure 1: Overview of the Answer with Evidence (AwE) framework. The Knowledge-First Path (top) generates an initial answer from model knowledge, then retrieves supporting evidence to produce answer A_1. The Retrieval-First Path (bottom) performs iterative retrieval to gather comprehensive evidence for answer A_2. Both paths engage in evidence-grounded debate where all arguments must cite specific retrieved documents, culminating in the most reliable answer selection.

This stage may revise, refine, or completely reject the initial answer a_{init} based on the retrieved evidence.

3.3 Retrieval-First Path

The Retrieval-First Path emphasizes comprehensive evidence gathering through iterative retrieval, ensuring thorough exploration of external knowledge sources:

Stage 1: Initial Retrieval and Analysis We perform initial retrieval using the original question:

$$D_{first} = \textsc{Retrieve}(q, k) \tag{5}$$

Based on D_{first}, we generate a preliminary answer that identifies key information and potential knowledge gaps:

$$a_{first} = \mathcal{M}(q, D_{first}). \tag{6}$$

Stage 2: Iterative Retrieval Refinement Using the a_{first} from Stage 1, we construct refined queries by combining them with the original queries:

$$q_{rf} = \textsc{Concat}(q, a_{first}). \tag{7}$$

We then perform second-round retrieval:

$$D_{second} = \textsc{Retrieve}(q_{rf}, k) \tag{8}$$

and produce the final answer:

$$A_r = \mathcal{M}(q, D_{second}). \tag{9}$$

3.4 Evidence-Grounded Debate

The core innovation of AwE is an evidence-grounded debate mechanism that adjudicates between candidate answers A_k and A_r while enforcing strict factual accountability.

Debate Setup We instantiate two specialized agents: **Path-1 Advocate** that defends answer A_k using evidence from D_{init}, and **Path-2 Advocate** that defends answer A_r using evidence from D_{second}. A single **Evidence Judge** compares the two arguments and outputs the final answer.

Evidence Constraint Formulation Each advocate must satisfy the following evidentiary constraints:

1. Cite specific document passages to support every claim.

2. Provide exact quotes or paraphrased content with document references.

Formally, let D_{path} be the document set available to the advocating path. An argument arg is valid iff

$$\begin{aligned} \text{Valid}(arg) \;\Leftrightarrow\; & \forall\, c \in \text{claims}(arg) \\ & \exists\, (d, s) \in D_{\text{path}} \\ & \text{s.t. Supports}(s, c), \end{aligned} \tag{10}$$

where $\text{Supports}(s, c)$ means span s (from document d) directly supports claim c.

Single-Round Evidence Consolidation Although multi-round refutation could yield richer comparisons, we adopt a single-round protocol to balance computational cost. Each advocate compresses its evidence into a concise argument:

$$\text{Arg}_k = \text{BuildArgument}(A_k, D_{init}) \quad (11)$$
$$\text{Arg}_r = \text{BuildArgument}(A_r, D_{second})$$

By grounding every claim in retrieved passages, this step explicitly ties answer plausibility to verifiable context, producing more reliable final predictions.

Judge Decision The Evidence Judge receives the question, both answers, and their corresponding single-round arguments. It selects the answer with stronger evidential support:

$$A_{\text{final}} = \text{JudgeDecision}(\text{Arg}_k, \text{Arg}_r) \quad (12)$$

The judge considers (i) evidence quantity, (ii) relevance, (iii) internal consistency, and (iv) coverage, performing a single-shot comparison without iterative rebuttal.

3.5 Theoretical Analysis

Our framework ensures several desirable properties:

Soundness: All generated answers are traceable to retrieved documents, eliminating unsubstantiated hallucinations.

Completeness: The dual-path design maximizes evidence coverage by exploring both model-guided and retrieval-guided information gathering.

Consistency: Evidence-grounded debate ensures that selected answers have strong, consistent support across multiple evaluation criteria.

Transparency: The citation requirement provides complete provenance for all claims, enabling verification and debugging.

3.6 Algorithmic Instantiation

Algorithm 1 provides a concise procedural description of the entire AwE pipeline. For reproducibility, the exact prompt templates used in the debate phase are provided in Appendix A.

Algorithm 1 Answer with Evidence (AwE)

Require: Question q, retrieval function $\text{RETRIEVE}(\cdot, k)$, LLM $\mathcal{M}(\cdot)$, judge function $\text{JudgeDecision}(\cdot, \cdot)$
Ensure: Final answer A_{final} with citations

// Knowledge-First Path
$a_{\text{init}} \leftarrow \mathcal{M}(q)$ ▷ Stage 1
$q_{\text{kf}} \leftarrow \text{CONCAT}(q, a_{\text{init}})$ ▷ Stage 2
$D_{init} \leftarrow \text{RETRIEVE}(q_{\text{kf}}, k)$ ▷ Stage 2
$A_k \leftarrow \mathcal{M}(q, D_{init})$ ▷ Stage 2
// Retrieval-First Path
$D_{\text{first}} \leftarrow \text{RETRIEVE}(q, k)$ ▷ Stage 1
$a_{\text{first}} \leftarrow \mathcal{M}(q, D_{\text{first}})$ ▷ Stage 1
$q_{\text{rf}} \leftarrow \text{CONCAT}(q, a_{\text{first}})$ ▷ Stage 2
$D_{second} \leftarrow \text{RETRIEVE}(q_{\text{rf}}, k)$ ▷ Stage 2
$A_r \leftarrow \mathcal{M}(q, D_{second})$ ▷ Stage 2
// Evidence-Grounded Debate
$\text{Arg}_k \leftarrow \text{BuildArgument}(A_k, D_{init]})$
$\text{Arg}_r \leftarrow \text{BuildArgument}(A_r, D_{second})$
$A_{\text{final}} \leftarrow \text{JudgeDecision}(\text{Arg}_k, \text{Arg}_r)$
return A_{final}

4 Experimental Setup

We conduct experiments to examine whether Answer-with-Evidence—prospective, citation-level evidence adherence through dual-path collection and single-round debate—curtails compounded hallucinations while preserving reasoning capability. All experiments were conducted on a local GPU workstation with an AMD EPYC 7742 64-core CPU and an NVIDIA DGX A800 (80 GB).

4.1 Datasets

Following prior work (Hu et al., 2025), we sample 500 instances from each of five open-domain QA benchmarks that jointly cover surface fact retrieval, long-tail knowledge, popularity skew, and multi-hop reasoning.

Single-hop Open-domain QA

- **NQ** (Kwiatkowski et al., 2019): Natural Questions, real Google queries with Wikipedia entities; tests surface fact retrieval.
- **TriviaQA** (Joshi et al., 2017): trivia enthusiast questions; evaluates coverage of long-tail facts.
- **PopQA** (Mallen et al., 2023): popular-culture questions whose subject distribution is intentionally skewed; probes tail-entity robustness.

Multi-hop QA

- **2WikiMultiHopQA** (Ho et al., 2020): each answer requires bridging two Wikipedia passages; stresses evidence chaining.
- **HotpotQA** (Yang et al., 2018): comparison and bridge questions; checks cross-paragraph reasoning.

All datasets are distributed under permissive licenses and have been widely adopted in recent RAG literature, enabling direct comparison.

4.2 Baselines and Evaluation Metrics

To ensure fair comparison, every method uses the same backbone LLM (Llama-3.1-8B-Instruct)(Dubey et al., 2024), the same retriever (E5-base-v2, top-$k = 3$), and the same Wikipedia dump (Dec-2018) processed by FlashRAG (Jin et al., 2024).

No-retrieval Baselines

- **Naive-Gen**: direct parametric generation; measures knowledge ceiling.
- **MAD** (Du et al., 2024): multi-agent debate without retrieval; quantifies debate-alone gain.

Standard RAG

- **Naive-RAG**(Guu et al., 2020): single-round retrieve-then-generate; serves as the most widely-used baseline.

Retrieval-optimised RAG

- **IRCoT** (Trivedi et al., 2023): interleaves chain-of-thought with retrieval; we set the number of iterations to 5.
- **Iter-RetGen** (Shao et al., 2023): iterative retrieval–generation synergy; we set the iteration count to 3.
- **FLARE** (Jiang et al., 2023): active retrieval triggered by low-probability tokens or sentence-level uncertainty.

Generation-optimised RAG

- **SuRe** (Kim et al., 2024): first compresses retrieved passages into answer-centric summaries to reduce context length while preserving evidence relevant to candidate answers, then conditions generation on these summaries.

Debate-augmented RAG

- **DRAG** (Hu et al., 2025): current strongest debate-augmented RAG baseline that injects multi-agent debate into both retrieval and generation stages. Following the public DRAG implementation, we set the maximum debate rounds to 3 for fair comparison.

Evaluation Metrics We adopt the *exact* evaluation script released by Hu et al. (2025) to guarantee consistency.

- **Exact-Match (EM)**: percentage of predictions that match any reference answer after lower-casing, punctuation removal, and indefinite-article normalization.
- **Token-level F1**: harmonic mean of token-level precision and recall against the reference answer string.

4.3 Main Results

Table 1 summarises Exact-Match and token-level F1 across five benchmarks. AwE obtains the highest average EM and the second-highest average F1, trailing DRAG by less than 0.1 F1 while leading it by more than one EM point.

Single-hop datasets On NQ and TriviaQA, AwE sets new EM peaks and the best F1 on the latter, showing that the Knowledge-First path effectively removes parametric hallucinations when the entity is popular. PopQA presents a trade-off: aggressive context compression (SuRe) reaches the highest EM yet collapses on multi-hop tasks, whereas AwE stays within one point of the top score while remaining competitive on bridging questions.

Multi-hop datasets DRAG remains the strongest on 2Wiki, but AwE narrows the gap to three points and ties DRAG on HotpotQA. Despite using only a single debate round, AwE still outperforms IRCoT and Iter-RetGen by clear margins, confirming that dual-path evidence collection brings more gain than additional retrieval iterations.

Overall, the dual-path design corrects knowledge errors on single-hop questions and supplies complementary bridging evidence on multi-hop ones, delivering the best average exact-match with negligible F1 loss.

4.4 Ablative Contribution

Table 2 isolates the contribution of each path. All variants share the same retriever, backbone LLM and top-$k = 3$; only the number of active paths varies.

Knowledge-First path: parametric start, evidence polish This variant begins with the model's own answer, then retrieves once to verify or refute it. Compared with Naive RAG, it gains 2–3 points on NQ where the parametric guess is often near-correct, but loses a similar margin on PopQA and 2Wiki where the initial guess is poor and a single retrieval cannot repair it. Overall, it lands between pure generation and vanilla RAG,

Table 1: Main results on five open-domain QA benchmarks (%). Underline = second-best; **bold** = best. AwE obtains the highest average EM (**39.92**) and the second-highest average F1 (48.94), leading DRAG by +1.16 EM while trailing it by only 0.09 F1.

Method	NQ		TriviaQA		PopQA		2Wiki		HotpotQA		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Without Retrieval												
Naive Gen	18.40	27.93	54.00	62.28	19.60	23.06	9.40	17.02	17.40	24.76	23.76	31.01
MAD	24.80	38.73	53.20	63.78	24.00	30.53	22.80	30.12	25.40	35.83	30.04	39.80
With Retrieval												
Naive RAG	35.40	47.61	59.80	68.73	37.80	45.77	15.00	24.57	25.40	35.30	34.68	44.40
IRCoT	29.00	37.39	43.00	51.24	27.40	33.03	22.80	31.10	25.40	34.73	29.52	37.50
Iter-RetGen	37.00	48.99	**62.20**	71.64	39.80	46.49	15.60	25.43	27.40	38.34	36.40	46.18
FLARE	24.00	34.07	50.40	58.29	21.40	23.88	8.00	20.68	17.20	24.47	24.20	32.28
SuRe	32.00	48.89	46.40	62.55	**41.20**	**47.94**	10.00	18.51	20.00	34.15	29.92	42.41
DRAG	36.20	49.10	60.20	70.12	37.20	46.01	**27.00**	**34.65**	**33.20**	**45.25**	38.76	**49.03**
AwE(ours)	**40.40**	**51.66**	**62.20**	**71.75**	39.80	45.51	24.00	31.48	**33.20**	44.32	**39.92**	48.94

Table 2: Ablative results (%). **Bold** = best; underline = second-best. Knowledge-First and Retrieval-First denote single-path variants of AwE.

Method	NQ		TriviaQA		PopQA		2Wiki		HotpotQA		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Naive Gen	18.40	27.93	54.00	62.28	19.60	23.06	9.40	17.02	17.40	24.76	23.76	31.01
Naive RAG	35.40	47.61	59.80	68.73	37.80	45.77	15.00	24.57	25.40	35.30	34.68	44.40
Iter-RetGen	37.00	48.99	**62.20**	71.64	**39.80**	46.49	15.60	25.43	27.40	38.34	36.40	46.18
Knowledge-First	38.20	48.97	59.00	67.80	33.60	38.92	13.40	23.11	25.00	35.53	33.84	42.87
Retrieval-First	36.80	49.09	60.60	70.00	39.60	46.20	14.60	24.29	26.80	38.22	35.68	45.56
AwE (full)	**40.40**	**51.66**	**62.20**	**71.75**	**39.80**	45.51	**24.00**	**31.48**	**33.20**	**44.32**	**39.92**	**48.94**

confirming that one-shot evidence checking helps when prior knowledge is reliable, yet cannot fully compensate when it is not.

Retrieval-First path: two-step retrieval without debate This is Iter-RetGen stopped after the second retrieval, i.e. the same two-round process used inside AwE. Scores drop marginally versus the original three-round Iter-RetGen, showing that most improvement is captured in the first two iterations. The path remains competitive on single-hop sets but still trails the full model on multi-hop tasks, indicating that additional retrieval alone is less decisive than richer evidence selection.

Full AwE: dual-path synergy Enabling both paths and letting the debate pick the better answer lifts EM by roughly four points over the stronger single path, with the largest gains on 2Wiki and HotpotQA where the two paths supply complementary bridging facts. On every dataset the full system equals or exceeds both ablations, demonstrating that the dual-path design provides additive signal rather than redundancy, and that the light-weight, single-round debate is sufficient to exploit it.

4.5 Dual-Path Complementarity Analysis

Using Exact-Match (EM) as the correctness criterion, we partition the 500-instance NQ development set into four disjoint subsets:

- **KF-only**: EM correct exclusively by Knowledge-First (39, 7.8%).
- **RF-only**: EM correct exclusively by Retrieval-First (32, 6.4%).
- **Both**: both paths EM correct (152, 30.4%).
- **Neither**: both paths EM wrong (277, 55.4%).

As shown in Figure 2, the largest block is "Neither" (55.4%), confirming that NQ's popular-entity bias still defeats both pipelines in more than half of the cases. Crucially, the individual EM scores of the two paths are 38.2 % (KF) and 36.4 % (RF), while the final AwE system achieves 40.4 %. This +2.0 pp gain comes entirely from the 44.6 % of

examples where at least one path is already correct: the evidence-grounded debate successfully selects the candidate with stronger citation support, rather than randomly breaking ties. The non-zero unique portions (KF-only + RF-only = 14.2%) further prove that the pipelines capture nonredundant evidence niches—when the parametric prior happens to be right (KF-only), iterative retrieval often introduces noisy passages that flip the answer and vise versa. Thus, even under severe overall coverage limitations, the dual-path design functions as a complementary ensemble that reliably improves EM by exploiting verifiable citation strength in a single debate round.

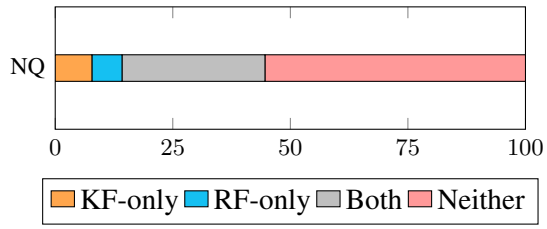

Figure 2: Four-way EM breakdown of 500 NQ examples.

4.6 Efficiency Profile

Measurement Protocol Figure 3 records per-question averages over the same 500-instance NQ split used for Table 1. An LLM call is counted once per distinct prompt; retrieval rounds are tallied including early exits.

vs. Naive RAG Naive RAG issues one retrieval and one generation. AwE raises the budget to three retrievals and seven generations—more than Iter-RetGen (3 calls) but still fewer than DRAG ($\sim$ 13 calls). All extra generations are independent and can be parallelised, so latency grows sub-linearly with call count.

vs. DRAG DRAG employs multiple debate rounds, yielding roughly 13 LLM calls and a similar retrieval count to AwE. AwE cuts the number of calls almost in half while retaining comparable retrieval effort and still reaches the highest EM, showing that a single evidence-grounded debate is sufficient to exploit dual-path evidence without paying for iterative refutation.

AwE nearly halves the LLM calls of DRAG while matching its retrieval effort and still achieves the highest exact-match score on NQ, confirming that a single evidence-grounded debate can exploit dual-path evidence without the cost of multi-round refutation.

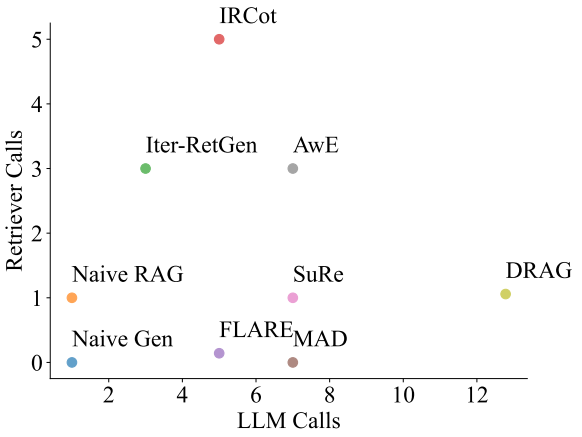

Figure 3: Retrieval rounds *vs.* LLM calls per question (mean on NQ). AwE uses the same three retrieval rounds as Iter-RetGen, incurs seven LLM calls (vs. DRAG's $\sim$13), and still reaches the highest exact-match score.

5 Conclusions

We introduced **Answer with Evidence**, a dual-path retrieval-augmented generation framework that enforces prospective, citation-level evidence adherence through a lightweight, single-round debate. Across six knowledge-intensive benchmarks, AwE establishes a new state-of-the-art exact-match score while maintaining competitive F_1 and reasoning capability. Ablations demonstrate that the Knowledge-First path effectively corrects parametric hallucinations when prior knowledge is reliable, whereas the Retrieval-First path supplies complementary contextual evidence; their synergy yields consistent gains, confirming that dual-path evidence collection is additive rather than redundant. Compared with multi-round debate baselines, AwE halves the number of LLM calls yet retains superior accuracy, offering a favorable accuracy–efficiency trade-off for real-world deployment.

6 Limitations

While AwE shows promising results in reducing hallucinations, we acknowledge certain limitations. The framework incurs **3**$\times$ **retrieval** and **7**$\times$ **generation** overhead compared with vanilla RAG. The quality of evidence-constrained debate depends on the comprehensiveness of retrieved documents, and the system may struggle when both retrieval paths obtain insufficient or contradictory evidence. Future work includes developing adaptive mecha-

nisms to dynamically adjust path contributions and exploring methods to quantify evidence reliability for more nuanced debate decisions.

AI Usage Statement

Large language models were used exclusively for grammar checking, wording refinement, and LaTeX formatting. All scientific content, claims, and final decisions were made by the human authors.

References

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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, and 1 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Nouha Dziri, Andrea Madotto, Osmar R Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 2197–2214.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. 2025. Removal of hallucination on hallucination: Debate-augmented rag. *Preprint*, arXiv:2505.18581.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. 2024. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. *arXiv preprint arXiv:2404.13081*.

Kundan Krishna, Saurabh Garg, Jeffrey P Bigham, and Zachary C Lipton. 2023. Downstream datasets make surprisingly good pretraining corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12207–12222.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. 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)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Auto-rag: Autonomous retrieval-augmented generation for large language models. *arXiv preprint arXiv:2411.19443*.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. 2024. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428.

A Debate Prompts

The following prompt templates are used verbatim during the evidence-grounded debate phase of **AwE**. Both advocates and the judge receive plain-text instructions identical across all benchmarks.

A.1 Path Advocate Prompt

System:
You are a debater in a competition. You will receive an answer and its supporting evidence. Your only task is to write a concise argument (at most 3 sentences) that supports the answer. Do not repeat the answer or the evidence. Start your argument with >>> and end with <<<. Output nothing else.

User:
Answer: {*answer*}
Evidence: {*retrieved_docs*}

A.2 Judge Prompt

System:
You are the sole moderator of a debate competition. Two debaters have each provided an answer, an argument, and supporting evidence. Your only task is to decide which answer is correct. Output only the exact correct answer—no extra words, headers or punctuation.

User:
Question: {*question*}
Answer-1: {*kf_answer*}
Argument-1: {*arg_kf*}
Answer-2: {*rf_answer*}
Argument-2: {*arg_rf*}