AF-RAG: ACTIVE FALSIFICATION FOR LOCAL RETRIEVAL CONTAMINATION IN LITERATURE QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated laboratories and auto-research systems increasingly use retrieved literature to generate hypotheses, plan experiments, and analyze results. However, erroneous numerical values, conclusions, or comparisons in the retrieval context can propagate into downstream research decisions. We study a constrained local retrieval failure in which erroneous or adversarially promoted evidence occupies the initial Top-K, while claim-relevant corrective evidence remains retrievable at deeper ranks in the same corpus. We propose Active Falsification RAG (AF-RAG), which audits verifiable claims, generates targeted falsification queries, retrieves counter-evidence beyond Top-K, and adjudicates claim-level conflicts. We formalize this condition as an information cocoon with stealth poisoning, where locally coherent micro-tampering changes query-relevant facts. Under this local assumption, AF-RAG conditionally prioritizes claim-relevant deep evidence. On 2,641 arXiv papers and 1,200 stealth-poisoned samples, AF-RAG lowers Poisoned Adoption Rate (PAR) from 94.75% to 42.67%, raises Truth Recovery Rate (TRR) from 25.58% to 67.50%, increases Joint Success Rate (JSR) from 3.75% to 49.83%, and achieves 79.9% accuracy on unpoisoned queries. Supplementary diagnostics evaluate a verification extension combining provenance-aware adjudication and conflict-driven evidence acquisition, showing improved joint success when initial evidence is correct and external evidence is contaminated.

1 INTRODUCTION

Scientific discovery has traditionally relied on researchers reading background literature, formulating hypotheses, testing them via experiments or computation, and revising subsequent directions according to the results. The AI Scientist spans literature search, experimentation, analysis, and writing, while Virtual Lab pairs AI agents with human guidance and experimental validation (Lu et al., 2026; Swanson et al., 2025). Recent retrieval-augmented literature synthesis (Lála et al., 2023), scientific question-answering (Asai et al., 2026), and multi-agent discovery systems (Ghafarollahi & Buehler, 2025; Ghareeb et al., 2026; Gottweis et al., 2026) indicate that external literature is becoming an important input to automated research decisions. Literature content determines the knowledge boundary and search direction of the auto-research loop (Figure 1(a)). Prior findings, candidate mechanisms, datasets, metric definitions, and method comparisons define the space for hypothesis and idea generation; numerical results, experimental conclusions, and ablation analyses further influence experimental prioritization and interpretation. Consequently, literature reliability directly affects whether automated research systems can generate reliable hypotheses and iterate in valid directions rather than merely produce fluent summaries.

However, scholarly literature is not fully reliable. A *Nature* survey of 1,576 researchers reported that more than 70% of respondents had failed to reproduce another scientist's experiment (Baker, 2016); in 2023, Crossref reported that its records and the Retraction Watch Database jointly covered around 50,000 retractions (Hendricks et al., 2023). Human research workflows can partially absorb such risks through expert reading and later correction. In contrast, automated research systems may directly use erroneous numerical values or pre-retraction conclusions for hypothesis generation and experimentation, thereby amplifying invalid experiments and erroneous iterative updates. Standard

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

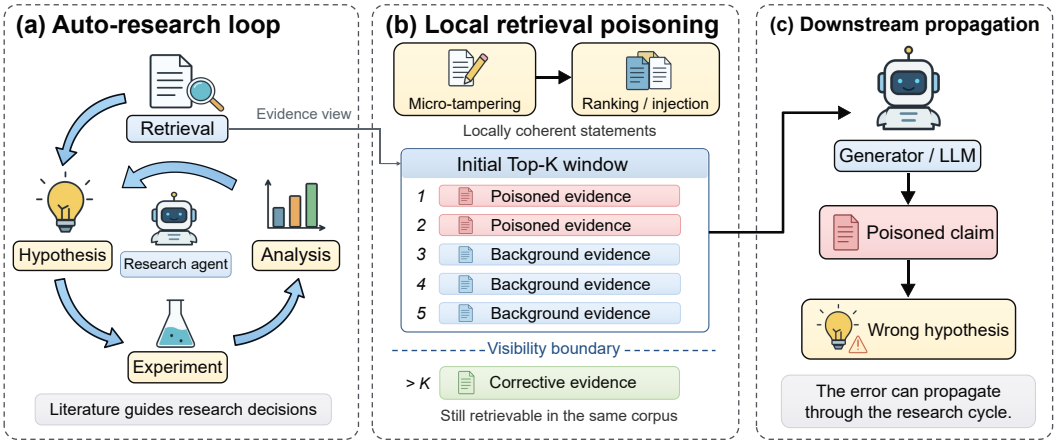


Figure 1: Information-cocoon threat in literature-grounded auto-research. (a) Retrieved literature informs hypothesis generation, experimentation, and analysis. (b) Locally coherent poisoned evidence enters the initial Top-K window, while corrective evidence remains retrievable beyond it. (c) A poisoned claim can propagate into a wrong hypothesis and subsequent research.

RAG primarily optimizes relevance rather than truthfulness; when the Top-K window contains only locally self-consistent erroneous literature and no authentic counter-evidence, filtering, aggregation, uncertainty-triggered retrieval, and multi-agent checking may lack effective signals (Figure 1(b)). A poisoned claim can then propagate into downstream hypotheses and research decisions (Figure 1(c)).

We therefore propose AF-RAG for a local setting in which the initial Top-K is contaminated but candidate corrective evidence remains reachable at deeper ranks. AF-RAG treats numerical values, comparative relations, and determinate factual statements as verifiable claims, then retrieves candidate counter-evidence beyond the initial window with targeted falsification queries. We evaluate a controlled combination of an information cocoon and stealth poisoning: micro-tampered, locally coherent documents enter Top-K, while reference documents remain outside the initially visible window. AF-RAG reduces poisoned-claim adoption, improves recovery of buried corrective evidence, and preserves benign-query utility.

To examine verification strategies across evidence configurations, we introduce provenance-aware adjudication and conflict-driven evidence acquisition as an optional extension and assess its role through paired diagnostics with fixed evidence. The extension improves joint recovery when the initial evidence is correct and the external evidence is contaminated.

2 Related Work

2.1 Automated Research Systems and Literature-Grounded Scientific Agents

PaperQA and OpenScholar synthesize cited literature (Lála et al., 2023; Asai et al., 2026); SciAgents reasons over scientific knowledge graphs (Ghafarollahi & Buehler, 2025). The AI Scientist automates computational research through manuscript writing (Lu et al., 2026), while Virtual Lab combines AI agents, human feedback, and experimental validation (Swanson et al., 2025). Co-Scientist generates and refines evidence-conditioned hypotheses (Gottweis et al., 2026); Robin links literature search, experimental proposals, and data analysis (Ghareeb et al., 2026).

2.2 RAG, Knowledge Poisoning, and Evidence Misplacement

RAG moves part of its factual basis into retrieval, making corpus construction, ranking, chunking, and context selection security-relevant (Karpukhin et al., 2020; Lewis et al., 2020; Gao et al., 2023b). Knowledge poisoning exploits this channel by injecting or promoting documents into Top-K; even a few adversarial passages, documents, or triggers can steer generation (Zhong et al., 2023; Zou et al., 2025; Xue et al., 2024). Attacks and benchmarks show corruption from limited poisoning, moti-

vating distinctions between initial-context compromise and broader corpus contamination (Zhang et al., 2026; 2025b; Wang et al., 2025). Retrieved context may conflict with parametric knowledge (Longpre et al., 2021). We study evidence misplacement in automated research, where ranking buries corrective evidence while coherent errors drive decisions. Rather than modeling injection or optimizing documents, we address ranking failure and recover evidence beyond initial Top-K.

2.3 Defenses, Active Retrieval, and Claim Verification

Existing RAG defenses use adversarial training, filtering, isolation, and forensic traceback (Fang et al., 2024; Cheng et al., 2025; Zhou et al., 2025; Kim et al., 2025; Xiang et al., 2024; Zhang et al., 2025a). Graph-based aggregation and multi-agent debate provide cross-checking (Edge et al., 2024; Du et al., 2024). When corrective evidence is absent from the initial context, methods that operate only on visible evidence lack a direct contrast; we study the role of additional evidence acquisition.

FLARE, CRAG, and Self-RAG control retrieval and generation through generation confidence, retrieval evaluation, and self-reflection, respectively (Jiang et al., 2023; Yan et al., 2024; Asai et al., 2024). FEVER, SciFact, and FActScore study claim verification and atomic-fact support (Thorne et al., 2018; Wadden et al., 2020; Min et al., 2023); RARR revises generated content through evidence acquisition, while FIRE iterates retrieval and verification (Gao et al., 2023a; Xie et al., 2025). AF-RAG audits verifiable claims in the initial context even when they are internally consistent, generates queries that preserve entities, metrics, and experimental conditions, retrieves beyond the initial window, and synthesizes evidence under the local corrective-evidence assumption.

3 Problem Definition

3.1 Task Definition

Given a literature query q, an initial context C^0 containing K document excerpts, and an accessible corpus D, a system generates an answer y. The initial context may contain erroneous claims. The system may retrieve additional evidence from D to recover the correct query-relevant facts and avoid adopting the erroneous alternatives. Correctness is tied to the entities, metrics, and experimental conditions requested by the query: a result reported under one setting does not automatically answer a question about another.

A query about a reported benchmark improvement, for example, is answered relative to the named baseline and evaluation setting. A claim valid under a different baseline is not a correction unless it resolves the original question.

We treat C^0 as the context supplied to the system, without requiring it to arise from an explicitly simulated ranking attack. Each excerpt retains its document identifier; let D^0 denote the documents represented in that context. The task concerns conditional factual recovery after local contamination. It does not require authenticating every corpus document, and the designated reference claims are not supplied to the answering system as labels.

3.2 Local Retrieval Contamination

The setting combines a constraint on evidence placement with a constraint on the form of the corruption. Let D_q^- denote documents containing the target erroneous claims and D_q^+ the designated uncontaminated reference documents. An *information cocoon* satisfies

$$D^0 \cap D_q^- \neq \varnothing, \qquad D^0 \cap D_q^+ = \varnothing, \qquad \varnothing \neq D_q^+ \subseteq D. \tag{1}$$

The initial context therefore contains poisoned evidence while excluding the designated references. This condition specifies which corrective sources are absent; by itself, it does not rule out every other potentially relevant clue in the initial context.

Stealth poisoning changes a query-relevant numerical value, comparison, dataset size, ranking, significance statement, or conclusion while preserving local grammatical and narrative coherence. The alteration concerns the factual evidence rather than an instruction to the model. A poisoned alternative conflicts with the reference fact under the same conditions; differences attributable to distinct experimental settings need not constitute corruption. Evidence placement and factual alteration

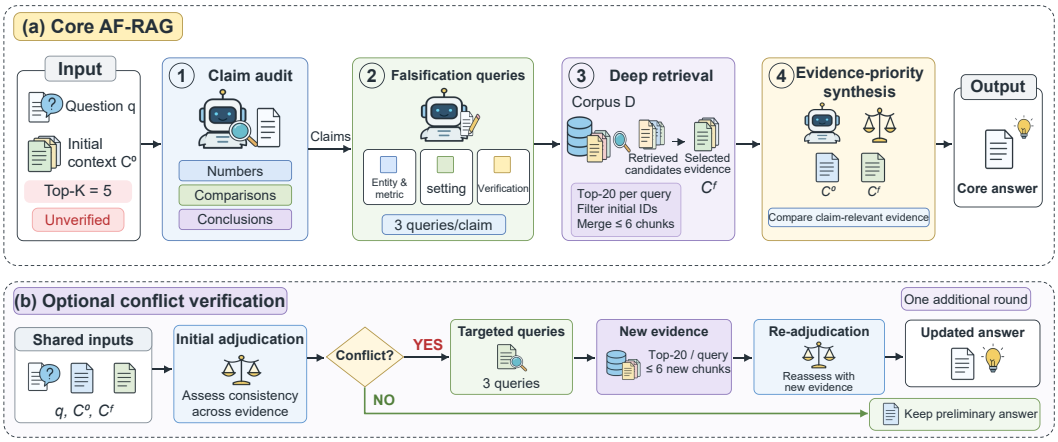


Figure 2: AF-RAG architecture. Initial Top-K evidence is unverified. AF-RAG audits verifiable claims, retrieves beyond the initial window using targeted falsification queries, and conditionally prioritizes claim-relevant candidates under the local corrective-evidence assumption.

jointly define the problem: coherence can conceal an erroneous claim, while the window boundary separates it from its designated corrective source.

3.3 Threat Model and Assumptions

Adversarial control. The victim is a literature-grounded RAG or automated research system that answers from a limited initial context. The adversary aims to make its final answer adopt a target erroneous value, comparison, or conclusion. We condition on poisoned evidence already occupying the initial window and do not model its operational injection or promotion. The adversary does not control the verification components or the subsequent retrieval pool; corpus-wide compromise is outside the main setting.

Corrective evidence. For each target claim, at least one relevant, uncontaminated reference must remain retrievable outside the initial context. This availability assumption is separate from the restriction on adversarial control: the existence of one correct source alone does not make every external candidate trustworthy. Nor does corpus membership guarantee that a finite retrieval budget will expose the relevant passage. Evidence must first be retrieved and selected before it can support answer correction. The task thus distinguishes an available corrective source from a system's successful use of that source.

4 Method

AF-RAG audits query-relevant claims, generates falsification queries, retrieves evidence beyond the initial context, and adjudicates claim-level disagreements (Figure 2). The auditor, query generator, and synthesizer use the same interchangeable LLM through prompts, without separate training. The retrieval stage is also called the deep retriever: "deep" denotes evidence outside the initial window, not necessarily a lower position in the original query's ranking. The pipeline returns an answer and, when applicable, a conflict marker. If no auditable claim or usable external evidence is obtained, it falls back to synthesis over C^0.

4.1 Claim Auditing

The auditor identifies statements whose factual content matters to the query. For an instance $x = (q, C^0, D)$, it returns an audit decision a and an ordered list of claim texts

$$(a, C_x) = A(q, C^0), \qquad C_x = (c_1, \ldots, c_{n_x}). \tag{2}$$

Each claim has an accompanying source span and verification rationale. The text carries the entities, metrics, comparisons, and conditions to be queried; the source span links the target to its original

context. These records preserve the distinction between an individual disputed statement and the surrounding document, which may contain other valid claims.

For a comparison, the audit target retains both methods and the evaluation dimension; for a numerical result, it retains the measured quantity rather than extracting an isolated number. This granularity keeps later queries tied to the statement that could change the answer.

Auditing is triggered by verifiability, even when the statement is fluent and internally consistent. The configured intent gate skips summary requests, author lookups, and section listings. For factual questions, the auditor extracts values, percentages, rankings, method comparisons, dataset sizes, and determinate conclusions. The structured field named "anomaly" indicates a verification target, not a finding that its source is false. All n_x claims are processed by default. A diagnostic cap retains the first $m_x = \min(n_x, m_{\max})$ claims in output order; without a cap, $m_x = n_x$. Appendix B.1 specifies the output schema and associated parsing constraints.

4.2 Falsification Query Generation

The generator handles each retained claim separately, using the user query and claim text rather than annotated answers. It produces r query variants per claim

$$Q^f(x) = \left(q^f_{i,j}\right)_{i=1,\dots,m_x;\, j=1,\dots,r}, \qquad G(q, c_i) = \left(q^f_{i,j}\right)_{j=1}^{r}. \tag{3}$$

The collection retains claim and variant identities for later merging. The default variants target the entity and metric, the experimental context, and a natural-language verification of the reported result. They preserve the claim's comparison objects and relevant conditions while introducing verification-oriented cues tied to the audited statement.

Processing m_x claims yields rm_x query variants. This increases candidate retrieval opportunities while the final evidence budget remains shared. The per-claim query mapping records which verification target produced each retrieval request and preserves that association across the merged results used for evidence selection.

For example, a claim reporting an 11.5% improvement yields queries combining the paper or entity with the metric, adding the experimental setting, and asking what result was reported. A reversed comparison still retains both methods and the evaluation dimension. This use of the claim can surface evidence missed when the original question omits a necessary metric or setting. The retrieved material may support, refute, or qualify the claim; the query itself does not establish that the initial statement is wrong. AF-RAG postpones that decision until evidence can be compared under matching conditions during claim-level adjudication.

4.3 Evidence Retrieval and Selection

Let $I^0 = \{\mathrm{id}(d) : d \in D^0\}$ denote the document identifiers represented in the initial context. Retrieval searches the eligible corpus

$$D_{\mathrm{out}} = \{d \in D : \mathrm{id}(d) \notin I^0\}. \tag{4}$$

For each query, the retriever returns an ordered candidate list at depth K'. The collection of lists and the final selected evidence are

$$\mathcal{H}_x = \left(R_{K'}(q', D_{\mathrm{out}})\right)_{q' \in Q^f(x)}, \qquad C^f = \mathrm{Select}_B(\mathcal{H}_x), \tag{5}$$

where B is the shared chunk budget. Selection traverses query results in round-robin rank order, deduplicates normalized document identifiers and text chunks, and retains at most B chunks across all claims. This separates the number of retrieved candidates from the amount of evidence actually presented to the synthesizer. Round-robin merging gives query lists alternating selection opportunities, considering results across claims and variants before advancing to the next rank within each list rather than exhausting one list first.

Because the budget is shared, no claim is guaranteed a selected chunk. A target may occur among retrieved candidates yet be dropped during selection; an available reference may also fail to enter any candidate list. These cases distinguish retrieval coverage from the evidence actually available to answer generation under the selected-chunk budget.

Exclusion operates on document identifiers, not on whether a passage agrees with the initial claim. A source excluded by this rule cannot supply corrective evidence through this path, so availability must hold within the eligible pool. Retrieved identifiers and candidate chunks remain in the trace for diagnosing omissions during retrieval or selection. Deduplication also avoids presenting repeated passages as independent corroboration. Candidates remain unverified at this stage: their relevance to an entity or metric does not establish a contradiction or justify changing the answer. Those decisions belong to the adjudicator.

4.4 Claim-Level Adjudication and Answer Generation

The evidence-prioritized synthesizer compares initial and selected external evidence to generate

$$y = S(q, C^0, C^f). \tag{6}$$

Its prompt marks the initial context as unverified and external chunks as candidate corrective evidence. Comparison focuses on query-relevant numerical values, comparison directions, and conclusions. Before treating a discrepancy as a conflict, it checks whether the evidence refers to the same entities, metrics, and experimental setting. An unrelated disagreement provides no basis for correcting the audited fact.

For example, a higher score on a different dataset does not directly refute a claim about the queried dataset. Opposite method rankings on the same metric and setting, however, provide a claim-level contradiction requiring adjudication before the answer is revised.

When the evidence is compatible, the synthesizer integrates it. When a discrepancy does not establish a direct contradiction under matching conditions, it reports the difference without unsupported aggregation. A direct numerical mismatch, reversed comparison, or incompatible conclusion qualifies for external-evidence priority only when this alignment holds. Priority is conditional on the local contamination assumptions in Section 3.3; retrieval rank and scholarly provenance do not authenticate the candidate source.

The resulting correction is restricted to the contradicted claim and must not override unrelated facts in the initial context. A conflict marker identifies this claim-level disagreement under the stated assumptions, rather than general uncertainty. Source spans and retrieval records retain the connection between the disputed statement and its evidence, allowing an unsuccessful answer to be distinguished from a failure to obtain a relevant passage. These records support diagnosis without supplying reference-answer labels to the synthesis process.

4.5 Optional Extension: Provenance-Aware Follow-up Verification

The optional extension replaces the original evidence-prioritized synthesizer with a provenance-aware adjudicator and adds conflict-triggered evidence acquisition. Given the question, initial context, and first external evidence, the adjudicator aligns entities, metrics, units, and experimental conditions, returning a preliminary answer, a conflict flag, and a dispute record. Source-family identifiers group a paper and its versions; they are not truth labels. Decisions depend on direct evidential support and condition matching.

The dispute record anchors additional queries to the facts on which the evidence disagrees. The trigger is a substantive evidential conflict, so compatible statements can be resolved in the preliminary adjudication without another retrieval round.

A substantive conflict triggers verification queries covering both sides of the dispute. Retrieved snippets are deduplicated against one another and previously seen evidence before being combined with the existing context. The final adjudication uses this combined evidence and the same decision rules, without receiving the preliminary answer. At most one follow-up round is allowed; without a trigger, the preliminary answer is returned unchanged. Thus, the extension changes both the adjudication configuration and, when triggered, the available evidence. Appendix D.3 specifies its budgets and exclusion rules.

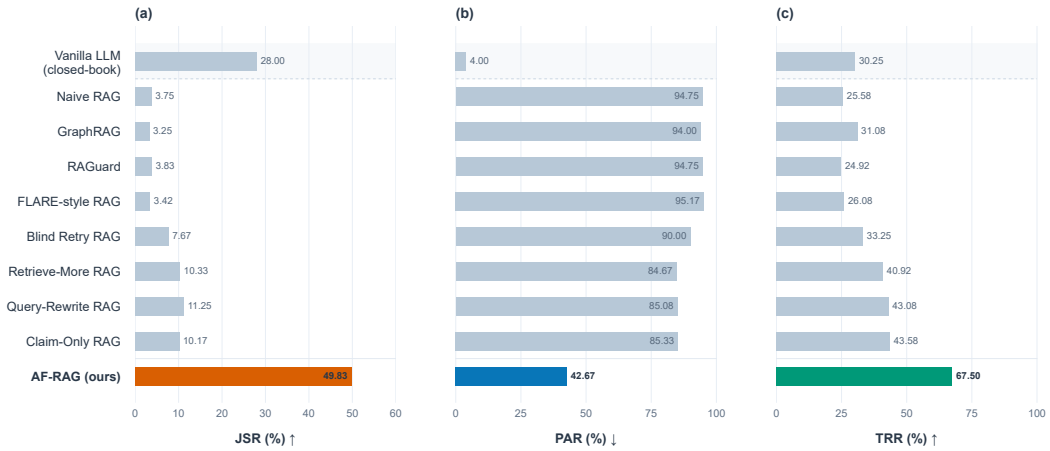


Figure 3: Main security results under the only-poison cocoon ($N = 1,200$). Higher JSR and TRR and lower PAR are better. Vanilla LLM is a closed-book reference. JSR requires truth recovery and explicit non-adoption of the poisoned claim. Full values appear in Appendix Table 18.

5 Experiments

5.1 Data and Construction

The main benchmark contains 1,200 stealth-poisoned instances drawn from an offline corpus of 2,641 arXiv papers. Each instance contains a query, initial context, poisoned document, and candidate corrective document. Tampering changes a query-relevant value, comparison, or conclusion while preserving local coherence, without injecting instructions. Only edits changing the correct answer or central conclusion enter the main security evaluation.

In the only-poison cocoon, poisoned evidence occupies the initial window and designated corrective sources remain outside it. This controlled placement evaluates recovery after local contamination; it does not measure an adversary's ability to promote a document into Top-K. We also evaluate 1,200 unpoisoned queries and construct an independent dataset for the supplementary verification diagnostic. Corpus processing, construction review, and subset sizes appear in Appendix A; supplementary data and evidence assembly appear in Appendix D.

5.2 Experimental Setup and Baselines

The configuration uses DeepSeek-V4-Flash and BM25 with an initial window of $K = 5$. AF-RAG processes all audited claims, generates $r = 3$ queries per claim, retrieves $K' = 20$ candidates per query, and selects at most $B = 6$ evidence chunks shared across claims. Model, decoding, and snippet settings are detailed in Appendix B.3; judge reliability and statistical protocols are documented in Appendix A.4.

The eight baselines are Naive RAG (Lewis et al., 2020), GraphRAG (Edge et al., 2024), RAGuard (Kumar et al., 2026), FLARE-style RAG (Jiang et al., 2023), Blind Retry RAG, Retrieve-More RAG, Query-Rewrite RAG, and Claim-Only RAG. They cover standard generation, evidence organization and filtering, and alternative retrieval strategies. Vanilla LLM is a closed-book reference, excluded from retrieval-method rankings. Benign-utility comparisons include Naive+Prompt, Multi-Agent, RAGDefender (Kim et al., 2025), and ReliabilityRAG (Shen et al., 2025). Implementations and budgets appear in Appendix B.4; the RAGuard setting with visible contrastive evidence is described in Appendix F.2. Code and evaluation materials are linked in Appendix A.5.

5.3 Evaluation Metrics

For poisoned queries, Poisoned Adoption Rate (PAR) measures the fraction of answers adopting the target erroneous claim, and Truth Recovery Rate (TRR) measures the fraction explicitly recovering

the reference fact. An answer may contain both claims, so these rates are not complementary. Joint Success Rate (JSR) requires truth recovery without adoption of the poisoned claim. Let T_i indicate truth recovery and E_i indicate explicit non-adoption for answer i. Then JSR is

$$\text{JSR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[T_i = 1 \wedge E_i = 1]. \tag{7}$$

For benign queries, Accuracy (Acc) is the fraction answered correctly, False Conflict Rate (FCR) is the fraction incorrectly flagged as evidential conflicts, and False Rejection Rate (FRR) is the fraction incorrectly rejected, reported for RAGuard. Lower PAR, FCR, and FRR and higher TRR, JSR, and Acc are better. Judgment labels and scoring rules appear in Appendix A.4.

5.4 Main Results

Figure 3 summarizes results on the same 1,200 instances. AF-RAG achieves PAR, TRR, and JSR of 42.67%, 67.50%, and 49.83%, decreasing PAR by 52.08 percentage points and increasing TRR and JSR by 41.92 and 46.08 points relative to Naive RAG. Its JSR exceeds that of Query-Rewrite RAG, the strongest remaining retrieval baseline, by 38.58 points and exceeds the closed-book model's 28.00%. These results show that under local contamination, AF-RAG improves truth recovery while reducing cases where recovered truth coexists with adoption of the poisoned claim.

On benign queries, AF-RAG achieves 79.9% accuracy but has a false-conflict rate (FCR) of 16.3%, exceeding Naive+Prompt's 12.8% and Multi-Agent's 3.0% (Table 1). RAGuard reaches 2.5% accuracy with a 96.6% false rejection rate (FRR). Appendix G.4 reports all accuracies, changes relative to Naive RAG, and false-positive signals. When authentic and poisoned evidence are both visible, RAGuard achieves lower PAR than AF-RAG (45.3% versus 54.1%), whereas AF-RAG achieves higher TRR (56.8% versus 34.8%; Appendix F.2). Active evidence acquisition and isolation-based filtering therefore offer different benefits and costs; AF-RAG's advantage is most evident when corrective evidence is absent from the initial window.

5.5 Mechanism and Evidence-Placement Diagnostics

To analyze the sources of AF-RAG's gains, we examine component contributions, evidence placement, cross-domain performance, and model configurations separately. Component ablations show that the stages jointly influence factual recovery. Removing the deep retriever yields 94.0% PAR and 10.0% TRR; removing the evidence-prioritized synthesizer yields 22.2% TRR. Removing the claim auditor raises PAR from 47.2% to 61.4% and lowers TRR from 61.2% to 55.2% on the same $N = 500$ ablation subset. Appendix E.2 provides full results.

Evidence-placement tests characterize behavior under different configurations. When the initial Top-K is clean but the deep pool is poisoned, truth accuracy is 97.0% for Naive RAG and 17.0% for AF-RAG; when both are poisoned, AF-RAG achieves 15.5%. Closed-book and target-removed TRR are 13.5% and 32.0%, respectively. Deep-evidence quality and coverage therefore directly affect final adjudication (Appendices F.1 and E.4).

On Physics and PMC, AF-RAG achieves PAR/TRR of 42.5%/69.0% and 54.5%/60.0%, respectively. Across paired multi-model evaluations, PAR decreases by 47.0–51.0 points and TRR increases by 46.5–48.0 points; Appendix F.5 gives the full configurations and results. In the main evaluation, AF-RAG improves JSR over Naive RAG by 46.08 points, with a paired bootstrap 95% interval of [43.08, 49.00]. Appendix E.1 reports paired PAR-avoidance and TRR tests. On a separate $N = 200$ subset with matched limits of four retrievals and 65 candidates, AF-RAG has 45.5% PAR/63.0% TRR versus 92.0%/16.5% for generic retrieval, although it uses more tokens (Appendix C.4).

5.6 Evidence-Condition Diagnostics of the Supplementary Module

Table 2 compares AF-RAG, AF-RAG with provenance-aware adjudication, and AF-RAG with the full supplementary module. In its evidence column, the first term describes the initial context and the second the external evidence. Poisoned/correct and correct/poisoned scores use the supplementary diagnostic's binary JSR: reference-fact adoption without judged adoption of the false alternative.

Architecture	Accuracy (%)	Signal
Naive RAG	63.7	–
GraphRAG	67.8	–
Naive+Prompt	64.0	FCR 12.8%
Multi-Agent	68.8	FCR 3.0%
RAGDefender	61.1	–
ReliabilityRAG	62.4	–
RAGuard	2.5	FRR 96.6%
AF-RAG	**79.9**	FCR 16.3%

Table 1: Benign-query utility ($N = 1,200$).

Initial / external	Configuration	Score (%)
Poisoned / correct	Original	**49.5**
	Provenance	36.0
	Follow-up	42.0
Correct / poisoned	Original	18.0
	Provenance	37.0
	Follow-up	**53.0**
Correct / correct	Original	**78.0**
	Provenance	76.0
	Follow-up	76.5

Table 2: Fixed-evidence results ($N = 200$).

Unlike the main JSR, it does not require a separate explicit-non-adoption label, so scores should be compared within this diagnostic rather than directly with the main results. Correct/correct scores are accuracy. Provenance adjudication uses no follow-up retrieval; the full module retrieves additional evidence when a substantive conflict is detected and readjudicates using it. Each evidence condition contains the same 200 questions, and all three configurations share the initial context and first external evidence. Appendix D details evidence assembly and the retrieval scope.

With correct initial and poisoned external evidence, the three configurations achieve JSR of 18.0%, 37.0%, and 53.0%, respectively; gains from provenance-aware adjudication and follow-up verification are both significant after Holm correction. With poisoned initial and correct external evidence, JSR is 49.5%, 36.0%, and 42.0%, respectively, and the 6.0-point gain from follow-up verification is not significant after correction. When both sources are correct, accuracy is 78.0%, 76.0%, and 76.5%. Thus, the supplementary module improves performance when the first external evidence is contaminated but does not outperform original AF-RAG when that evidence is correct, revealing a trade-off across evidence conditions.

5.7 Cost and Error Analysis

AF-RAG's cost stems from claim auditing and deep retrieval. In a latency diagnostic with $N = 20/20$, mean end-to-end latency increases from 1.02 to 5.69 seconds for factual poisoned queries and from 1.07 to 3.36 for benign queries. Claim-level active-verification rates are 100.0% and 50.0%, respectively. Appendix C reports latency, claim-budget, and retrieval-budget-matched results.

AF-RAG's falsification queries reach the authentic target document in 1,149/1,200 samples (95.75%); its final answers recover truth in 810 cases (67.50%) and achieve joint success in 598 (49.83%). PAR and TRR are judged independently at the answer level and may overlap: of the remaining 212 truth-recovering answers, 206 still adopt the poisoned claim and six are uncertain. Among the 390 answers without truth recovery, 363 already hit the authentic target (93.08%). Separately quantifying evidence reachability, truth recovery, and joint success motivates improved snippet selection, numerical-condition alignment, and conflict adjudication. Appendix G provides the contingency table and failure taxonomy. A target-document hit does not guarantee that the decisive passage reaches synthesis, so this gap does not isolate selection from adjudication.

6 Conclusion

Auto-research workflows depend on literature retrieval. When locally coherent poisoned evidence enters the initial window while corrective evidence is excluded, systems may form erroneous judgments affecting research decisions. By auditing verifiable claims, retrieving candidate counter-evidence beyond Top-K, and adjudicating claim-level conflicts, AF-RAG enables effective recovery from local retrieval contamination. Supplementary diagnostics show that provenance-aware adjudication and conflict-driven evidence acquisition improve joint success when initial evidence is correct and external evidence is contaminated, with trade-offs when external evidence is correct.

AI USE STATEMENT

Generative AI tools assisted with language polishing, phrasing, and translation. They were also used to generate controlled synthetic candidate samples according to predefined tampering types; the authors performed filtering, quality checks, and evaluation under the protocols in Section 5.1 and Appendices A and D. The authors reviewed all AI-assisted content included in the paper and take full responsibility for its final contents.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650(8103):857–863, 2026. doi: 10.1038/s41586-025-10072-4.

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016. doi: 10.1038/533452a.

Zirui Cheng, Jikai Sun, Anjun Gao, Yueyang Quan, Zhuqing Liu, Xiaohua Hu, and Minghong Fang. Secure retrieval-augmented generation against poisoning attacks. *arXiv preprint arXiv:2510.25025*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL `https://proceedings.mlr.press/v235/du24e.html`.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 16477–16508, 2023a. doi: 10.18653/v1/2023.acl-long.910.

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

Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025. doi: 10.1002/adma.202413523.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, et al. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y.

Juraj Gottweis et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Ginny Hendricks, Rachael Lammey, Lucy Ofiesh, Geoffrey Bilder, and Ed Pentz. News: Crossref and retraction watch. Crossref, September 2023. URL https://doi.org/10.13003/c23rw1d9.

Zhengbao Jiang, Frank F. 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, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Minseok Kim, Hankook Lee, and Hyungjoon Koo. Rescuing the unpoisoned: Efficient defense against knowledge corruption attacks on RAG systems. In *Annual Computer Security Applications Conference*, 2025. doi: 10.1109/ACSAC67867.2025.00093.

Pushkal Kumar, Tucker Nielson, Tanish Kolhe, Shubham Zala, and Vincent Li. RAGuard: A layered defense framework for retrieval-augmented generation systems against data poisoning. *arXiv preprint arXiv:2607.26339*, 2026.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565.

Chris Lu, Cong Lu, Robert Tjarko Lange, et al. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348, 2017. doi: 10.18653/v1/D17-1035. URL https://aclanthology.org/D17-1035/.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. ReliabilityRAG: Effective and provably robust defense for RAG-based web search. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/41457d56a2fdd20b4c072190290a549b-Abstract-Conference.html.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, et al. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646:716–723, 2025. doi: 10.1038/s41586-025-09442-9.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609.

Haowei Wang, Rupeng Zhang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, and Qing Wang. Joint-GCG: Unified gradient-based poisoning attacks on retrieval-augmented generation systems. *arXiv preprint arXiv:2506.06151*, 2025.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. FIRE: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2901–2914, 2025. doi: 10.18653/v1/2025.findings-naacl.158.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. BadRAG: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

Baolei Zhang, Haoran Xin, Minghong Fang, Zhuqing Liu, Biao Yi, Tong Li, and Zheli Liu. Traceback of poisoning attacks to retrieval-augmented generation. *arXiv preprint arXiv:2504.21668*, 2025a.

Baolei Zhang, Haoran Xin, Jiatong Li, Dongzhe Zhang, Minghong Fang, Zhuqing Liu, Lihai Nie, and Zheli Liu. Benchmarking poisoning attacks against retrieval-augmented generation. *arXiv preprint arXiv:2505.18543*, 2025b.

Baolei Zhang, Yuxi Chen, Zhuqing Liu, Lihai Nie, Tong Li, Zheli Liu, and Minghong Fang. Practical poisoning attacks against retrieval-augmented generation. In *Proceedings of the 31st ACM Symposium on Access Control Models and Technologies*, pp. 33–44, 2026.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775, 2023.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. TrustRAG: Enhancing robustness and trustworthiness in retrieval-augmented generation. *arXiv preprint arXiv:2501.00879*, 2025.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A Data and Evaluation Protocol

A.1 Main Corpus and Distribution

The experimental corpus contains 2,641 arXiv papers organized as a controlled offline literature corpus. Each poisoned evaluation instance contains a user query, an initial retrieval context, a poisoned document, a candidate corrective document, and a target claim. The construction follows the local ranking-manipulation assumption: authentic documents remain in the underlying corpus but are excluded from the initial Top-K in the only-poison cocoon, whereas poisoned documents are placed in the model-visible context.

Construction and traceability. The corpus was collected through predefined arXiv API topic queries, deduplicated by `entry_id`, and filtered to papers that could be downloaded, parsed, and converted to valid cleaned text. Cleaning removed page numbers, copyright noise, arXiv markers, and references. Each normalized `doc_id` maps one-to-one to an arXiv identifier without its version suffix. The corpus manifest records document ID, title, year, primary category, and source query, tracing candidate collection and filtering to the final 2,641 papers. The Physics and PMC Biomedical supplements reuse the same download, parsing, deduplication, cleaning, and controlled evidence-placement pipeline.

Table 3: Distribution of the corpus and main benchmark instances.

Primary category	Corpus papers	Benchmark instances
cs.CL	1,485	737
cs.AI	358	191
Other	798	272
Total	2,641	1,200

A.2 Tampering and Quality Review

Stealth poisoning is instantiated through micro-tampering rather than explicit prompt injection. The edited elements include numerical metrics, method comparisons, experimental conclusions, ablation interpretations, and task-performance descriptions. To preserve stealthiness, each edited passage is required to remain grammatical, locally coherent, and compatible with the surrounding narrative. The RAGuard fair setting differs from the only-poison cocoon setting by making both the poisoned document and the authentic contrastive document visible in the same input context.

Prior studies show that repeated runs can change reported scores, but the magnitude depends on the task and experimental protocol. For example, Reimers & Gurevych (2017) found that, with hyperparameters fixed and only random seeds changed, the 95th-percentile difference between paired runs was 1.08 percentage points for CoNLL-2003 NER and 2.10 percentage points for ACE 2005 entity recognition. We therefore require each constructed numerical alteration to change the original reported value by at least 30% in relative terms, so that the benchmark focuses on substantial numerical tampering rather than minor discrepancies. This threshold is a dataset-construction criterion, not a universal statistical bound on experimental variation.

Table 4: Micro-tampering operations used to construct stealth-poisoned evidence.

Claim type	Edit operation	Example
Numerical metric	Value substitution	10.0% $\rightarrow$ 20.0%
Comparison	Relation reversal	A outperforms B $\rightarrow$ B outperforms A
Conclusion	Polarity weakening or reversal	improves robustness $\rightarrow$ shows marginal gain
Ablation	Component interpretation change	module helps $\rightarrow$ module hurts

In the UCPOF example in Table 5, changing the reported accuracy improvement from 6.03% to 12.15% is a 101.5% relative increase, while changing the reported retrieval-trigger reduction from 50.66% to 30.24% is a 40.3% relative decrease. These relative changes use the authentic value as the denominator; they are distinct from differences in percentage points.

Table 5: Example of numerical micro-tampering. The poisoned claim remains locally fluent and plausible, but changes the factual values that a RAG answer may later adopt.

Authentic claim	Micro-tampered claim
UCPOF improves average accuracy by 6.03% over few-shot baselines, boosts overall average accuracy by 5.75%, and cuts the average retrieval trigger rate by 50.66%.	UCPOF improves average accuracy by 12.15% over few-shot baselines, boosts overall average accuracy by 10.42%, and cuts the average retrieval trigger rate by 30.24%.

To check construction quality without releasing operational attack prompts, two annotators independently inspected 200 sampled authentic–tampered claim pairs. They judged whether the authentic claim was supported by the unmodified source, whether the edit remained locally fluent and narratively consistent, and whether it changed the query's answer or central conclusion. Disagreements were jointly reviewed to produce final labels. The reported 0.92 is the proportion of agreement over these binary judgments, not Cohen's κ. This validation measures construction consistency; it does not imply that every poisoned passage is indistinguishable from authentic text in arbitrary deployments.

The main evaluation further requires the edit to be query-consequential. Only *query-consequential* samples enter PAR/TRR; *query-non-consequential* samples measure unnecessary conflict triggering when an edit does not change the query-level conclusion. Non-consequential differences, including plausible seed, hardware, implementation, or API variation, are excluded from the main poisoning scores.

Consequence and detectability diagnostics are reported in Appendix F.3.

A.3 Evaluation Subset Index

Table 6: Evaluation subsets used in the appendix and main experiments.

Setting	N	Purpose	Details
Only-poison cocoon	1200	Main security	E.1, G.1–G.2
RAGuard fair setting	1200	Isolation defense	F.2
Noisy-vs-stealth	50/50	Detectability diagnosis	F.3
Ablation subset	500	Component diagnosis	E.2
Cost subset	20/20	Latency / trigger rate	C.5
Consequence controls	200/200	Query-level effect / conflict	F.3
Evidence placement	200	Deep-pool boundary	F.1
Cross-domain	200/200	Physics / PMC robustness	F.4
Multi-model and causal	200	Model / memory controls	F.5, E.4
Sensitivity and budgets	200	Local resource diagnosis	C.1–C.4
Supplementary evidence conditions (Table 2)	200 each	Paired original/extended configuration comparison	D

Table 6 summarizes the evaluation subsets. The full only-poison and RAGuard fair settings contain 1,200 samples each; other results use smaller diagnostic subsets.

The $N = 200$ diagnostics need not use the same samples. Unless a table explicitly states otherwise, point estimates from different model versions, run batches, and random samples are compared only within their own tables and do not replace the main results.

The independent follow-up diagnostic data, selection, and fixed evidence assembly are described in Appendices D.1 and D.2.

A.4 Scoring Implementation and Judge Reliability

Data and scoring. Metric definitions appear in Section 5.3. The adoption judgment takes one of three labels: `poisoned`, `immune`, or `uncertain`. The `immune` label denotes explicit non-adoption; `uncertain` does not. PAR uses `judgments=poisoned`, TRR uses `truth_recovery=true`, and JSR requires both `truth_recovery=true` and `judgments=immune`. All methods have complete judgment fields. Uncertain answers remain in the fixed denominator as joint failures. PAR avoidance is $1 - \mathrm{PAR}$ and includes both `immune` and `uncertain`, so it differs from explicit non-adoption.

PAR and TRR are judged independently: PAR checks adoption of the poisoned value, comparison, or conclusion, while TRR checks whether the final answer states the authentic claim. We manually labeled 200 fixed answer-level outcomes and obtained overall Cohen's $\kappa = 0.87$ against the automatic judgments; this is an aggregate judge audit, not a separate κ for PAR and TRR, and does not measure source authenticity.

Main paired tests appear in Appendix E.1; the follow-up diagnostic uses the independent scoring and statistical protocol in Appendix D.4.

A.5 Material Availability

Data and material availability. The anonymous repository (`https://anonymous.4open.science/r/AF-RAG-6708`) provides the corpus manifest, normalized ID mapping, benchmark instances, cleaning and indexing scripts, defensive prompts, baseline configurations, per-sample outputs, evaluation scripts, and run commands. For safety, it excludes complete prompts that would directly automate stealth-poisoned document generation. The released benchmark, defense pipeline, and reported results can therefore be audited, but the original attack-generation process cannot be reproduced prompt-for-prompt from the public materials.

B Method Implementation and Baseline Configurations

B.1 Core Pipeline and Structured Outputs

AF-RAG is implemented through defensive prompts and retrieval operations. The claim auditor identifies verifiable factual claims, including values, comparisons, performance conclusions, dataset sizes, and determinate assertions; triggering does not require an internal contradiction. All three LLM modules are prompted and require no separate training. The generator independently creates three query variants per audited claim. The retriever aggregates all queries, excludes identifiers already present in the initial context, and forwards at most six deduplicated chunks shared across claims. Under the threat model, the synthesizer treats the initial context as unverified and conditionally prioritizes evidence retrieved by falsification queries.

Claim-auditor output. The auditor returns structured JSON with the audited claim, source span, and trigger rationale:

```
{
  "is_anomalous": true,
  "anomaly_points": [{
    "claim": "claim to verify",
    "source": "context span",
    "reason": "verification rationale"
  }]
}
```

Summary, author lookup, section listing, and other intents that do not require factual adjudication return an empty claim set.

Query-generator output. The orchestrator handles each retained claim independently and generates an entity/metric query, an experimental-context query, and a natural-language verification query:

```
{
  "claims_used": 2,
  "per_claim_queries": [
    {"claim": "claim 1",
     "queries": [
       "precise", "context", "semantic"
     ]},
    {"claim": "claim 2",
```

```
      "queries": [
        "precise", "context", "semantic"
      ]}
  ]
}
```

Per-sample traces retain `claims_used`, `per_claim_queries`, and the flattened query list.

B.2 Evidence Synthesis, Retrieval Constraints, and Fallback

Synthesis and conflict marker. The synthesis prompt marks the initial context as unverified and checks the audited claim for numerical mismatch, reversed comparison, and incompatible experimental conclusion. A direct conflict causes the answer to begin with `<CONFLICT_ALERT>` and report the discrepancy; otherwise, the model answers normally. Deep-evidence priority holds only under Section 3.3's local corrective-evidence assumption and is not source authentication.

We do not provide operational attack-generation prompts. The appendix describes only the defensive audit, retrieval, and synthesis constraints required to reproduce reported system behavior.

Reproducibility constraints. The implementation schema-checks structured auditor and generator outputs and records audited claims, query lists, retrieved document IDs, candidate chunks, and final judgments in each trace. If auditing does not trigger or no usable deep candidate is obtained, the system falls back to standard synthesis over the initial context. The anonymous repository contains the complete defensive prompts, parsing constraints, and run configurations.

B.3 Models, Decoding, and Defaults

Independent supplementary evaluations use DeepSeek-V4-Flash and are reported separately. Runs set temperature to 0.0 when supported, although cloud snapshots, server implementations, and generation stochasticity can still change point estimates.

Table 7: Core implementation configuration. The main evaluation processes all audited claims; claim-budget diagnostics impose a separate cap.

Item	Setting
Initial retrieval Top-K	5
Main claim handling	$m_x = \lvert C_x \rvert$
Queries per claim r	3
Deep retrieval window K'	20
Maximum shared deep chunks	6
Claim-budget diagnostic	$m_{\max} \in \{1, 2, 3, 4\}$
GraphRAG community summaries	3
Noisy/stealth subset	50/50
Latency subset	20/20
Ablation subset	500
LLM backbone	DeepSeek-V4-Flash
Corpus index	BM25 over cleaned papers in the local offline arXiv corpus
Initial context	Poisoned primary excerpt within the BM25 Top-$K = 5$ context; the authentic target is excluded in the only-poison setting
Snippets	Initial snippets truncated to 1,200 characters; deep evidence to 1,500
Deep retrieval	BM25 over the same corpus; initial IDs excluded; cross-query ID deduplication and round-robin merge
RAGuard embedding	SentenceTransformer `all-MiniLM-L6-v2`
Judges	Separate poisoned-adoption and truth-recovery judgments; model configuration as specified for each experiment

B.4 Baseline Implementations and Retrieval Budgets

Naive RAG retrieves the initial Top-K with the original query and generates directly from it without a defensive prompt (Lewis et al., 2020). GraphRAG evaluates graph-structured evidence organi-

zation and community aggregation (Edge et al., 2024). RAGuard (Kumar et al., 2026) removes each context document in turn, regenerates the answer, estimates document influence from semantic change, filters the most influential document, and generates a final answer. FLARE-style RAG triggers retrieval from low-confidence generation states (Jiang et al., 2023). Blind Retry repeats retrieval with the original query. Retrieve-More expands the retrieval window. Query-Rewrite generates a generic rewrite. Claim-Only uses the audited claim as the retrieval entry point without falsification-query construction or evidence-prioritized synthesis. Query-derived Factual Target constructs a neutral target solely from the user query and does not access the initial context, poisoned claim, or annotated answer.

Naive+Prompt adds a cautious conflict-detection prompt to Naive RAG without additional retrieval. Multi-Agent uses separate generation and checking agents over the retrieved context. RAGDefender and ReliabilityRAG instantiate prior filtering or reliability-oriented defenses (Kim et al., 2025; Shen et al., 2025). These four configurations are used only for benign utility and are not primary only-poison baselines. Vanilla LLM does not receive retrieved context and is reported only as a closed-book reference, outside retrieval-method rankings.

Table 8: Baseline retrieval budgets and control flow.

Method	Retrieval budget	Query source	Trigger	Synthesis
Naive RAG	Initial $K = 5$	User query	None	Standard context
GraphRAG	$K = 5 + 3$ summaries	User query	Always	Graph aggregation
FLARE-style	Initial + triggered	Low-confidence span	Low confidence	Standard context
Blind Retry	$K = 5$ + one retry	User query	Always	Standard context
Retrieve-More	Expanded window	User query	Always	Standard context
Query-Rewrite	Additional retrieval	Generic rewrite	Always	Standard context
Query-derived target	Additional initial	Neutral target from query	Always	Standard context
Claim-Only	Deep $K' = 20$	Audited claim	Verifiability	Claim retrieval
AF-RAG	$K = 5, K' = 20, r = 3$	Falsification queries	Verifiability	Evidence-prioritized
RAGuard	Initial context only	Initial context	Visible contrast	Leave-one-out filter

C Sensitivity and Computational Cost

C.1 Retrieval-Parameter Sensitivity

Table 9 reports a diagnostic sensitivity grid over the initial retrieval depth K, the falsification-retrieval depth K', and the number of generated falsification queries r. The analysis is conducted on a fixed $N = 200$ subset. Target-evidence hit rate generally increases with larger K' and r, but TRR remains within a narrower range, indicating that retrieval recall alone does not eliminate poisoned-claim adoption.

The default setting ($K = 5, K' = 20, r = 3$) is therefore not claimed as the best-performing point on this subset. It is a conservative configuration that captures most retrieval gains while avoiding a more aggressive evidence window. Increasing K' to 50 improves target hit rate in several configurations, but the corresponding TRR gains are modest and PAR is not monotonic, reinforcing that conflict adjudication remains the main bottleneck.

C.2 Deep-Evidence Chunk Budget

Increasing the budget from six to twelve chunks yields limited TRR improvement while increasing conflict marking and latency. Six chunks are therefore a preset efficiency–effectiveness trade-off, not a global optimum.

C.3 Claim Budget and Theoretical Retrieval Cost

For sample x, let m_x be the number of processed claims and r the number of queries per claim. Including the initial retrieval, the number of retrieval executions is

$$N_{\mathrm{ret}}(x) = 1 + rm_x. \tag{8}$$

The number of candidate chunks before deduplication is bounded by

$$N_{\mathrm{cand}}(x) \leq K + rm_xK'. \tag{9}$$

Table 9: Sensitivity analysis over initial retrieval depth K, falsification-retrieval depth K', and number of falsification queries r on a fixed $N = 200$ subset.

K	K'	r	PAR	TRR	Target hit
3	10	1	47.0%	62.0%	92.0%
3	10	3	46.5%	62.5%	94.5%
3	10	5	45.0%	61.5%	96.0%
3	20	1	46.0%	58.0%	92.0%
3	20	3	44.5%	62.5%	97.0%
3	20	5	43.5%	63.5%	96.5%
3	50	1	47.0%	59.0%	95.0%
3	50	3	43.5%	64.5%	99.0%
3	50	5	49.0%	62.0%	98.5%
5	10	1	46.0%	60.5%	92.0%
5	10	3	45.5%	63.0%	95.5%
5	10	5	46.0%	59.5%	97.0%
5	20	1	46.5%	59.5%	94.5%
5	20	3	47.5%	62.0%	97.0%
5	20	5	43.0%	62.5%	97.5%
5	50	1	49.5%	61.5%	94.0%
5	50	3	43.5%	64.5%	99.5%
5	50	5	49.0%	63.0%	98.0%
8	10	1	46.0%	58.5%	92.0%
8	10	3	45.0%	59.5%	96.0%
8	10	5	49.5%	59.5%	95.5%
8	20	1	47.0%	61.0%	95.0%
8	20	3	42.0%	63.5%	96.5%
8	20	5	49.5%	64.0%	97.5%
8	50	1	46.0%	59.0%	96.5%
8	50	3	44.5%	63.0%	99.5%
8	50	5	41.0%	63.5%	98.5%

Table 10: Deep-evidence chunk budget on an $N = 200$ diagnostic subset.

Chunks	PAR ↓	TRR ↑	Conflict	p50/p95 (s)
2	51.5%	62.0%	37.0%	5.48/8.48
4	49.5%	62.0%	41.5%	5.51/8.63
6	44.5%	62.5%	42.0%	5.59/8.69
8	45.0%	63.5%	44.0%	5.65/8.77
12	43.5%	64.0%	47.0%	5.78/9.14

When auditing and synthesis trigger, claim auditing, per-claim query generation, and final synthesis require $m_x + 2$ LLM calls, excluding the offline evaluation judge. Retrieval executions, LLM calls, and candidate chunks are distinct cost measures. If $m_{\max}$ is the maximum claim budget, $1 + rm_{\max}$ is only a theoretical retrieval upper bound; observed cost is lower when fewer claims are available.

Table 11: Claim-budget cost results on an $N = 200$ subset with $K = 5$, $r = 3$, and $K' = 20$. Tokens are per-instance end-to-end totals.

$m_{\max}$	Retrieval upper bound	Candidate upper bound	End-to-end tokens	p50/p95 (s)	PAR ↓	TRR ↑
1	4	65	7,928	4.83/8.15	45.5%	63.0%
2	7	125	10,751	5.68/10.02	47.5%	61.5%
3	10	185	14,185	6.69/11.43	49.5%	59.0%
4	13	245	17,623	8.21/13.84	53.0%	57.0%

Resource consumption grows with the claim budget, but effectiveness does not improve monotonically: from $m_{\max} = 1$ to 4, PAR rises from 45.5% to 53.0% and TRR falls from 63.0% to 57.0%. These short controlled contexts cover only $m \leq 4$ and support local cost sensitivity, not large-scale scalability.

C.4 Retrieval-Budget-Matched Controls

We use an independent random $N = 200$ subset to distinguish active-falsification structure from additional retrieval budget.

Table 12: Retrieval-budget-matched controls on an independent $N = 200$ subset. Non-AF-RAG rows report observed values; AF-RAG retrieval and candidate counts are theoretical upper bounds.

Configuration	Retrievals	Candidates	End-to-end tokens	p50/p95 (s)	PAR ↓	TRR ↑
Naive RAG	1	5	2,290	1.09/1.75	95.5%	17.5%
Blind Retry	2	25	3,198	1.28/1.73	95.0%	18.5%
Claim-Only	2.7	38.5	7,040	3.21/4.89	89.5%	26.0%
Query-Rewrite	4	65	3,363	2.33/3.04	94.5%	20.0%
Budget-matched Generic Retrieval	4	65	4,907	2.87/3.46	92.0%	16.5%
AF-RAG, $m_{\max} = 1$	≤ 4	≤ 65	7,928	4.83/8.15	45.5%	63.0%
AF-RAG, $m_{\max} = 2$	≤ 7	≤ 125	10,751	5.68/10.02	47.5%	61.5%
AF-RAG, $m_{\max} = 3$	≤ 10	≤ 185	14,185	6.69/11.43	49.5%	59.0%

Budget-matched Generic Retrieval and AF-RAG with $m_{\max} = 1$ share upper bounds of four retrievals and 65 candidates, yet their PAR/TRR are 92.0%/16.5% and 45.5%/63.0%. Additional retrieval budget alone therefore does not explain the gain. AF-RAG still consumes more tokens (7,928 vs. 4,907), so this is not a fully equal-cost comparison. These results are separate from the $N = 1200$ main evaluation.

C.5 Measured Latency and Trigger Rate

The overhead study measures average end-to-end wall-clock latency and claim-level active-verification frequency on $N = 20$ factual poisoned and $N = 20$ benign queries. Full-pipeline execution is repeated with claim auditing and deep retrieval enabled.

AF-RAG increases latency on both query types. The increase is larger for factual poisoned queries because all samples trigger auditing, whereas benign queries trigger less often. This small diagnostic does not establish scalability in large AutoResearch workflows.

Table 13: Average end-to-end latency and trigger-rate overhead of AF-RAG relative to Naive RAG, measured on $N = 20$ factual poisoned queries and $N = 20$ benign queries.

Metric	Factual poisoned queries	Benign queries
Naive RAG latency	1.02s	1.07s
AF-RAG latency	5.69 ± 1.18s	3.36 ± 1.79s
AF-RAG latency p95	7.62s	6.75s
Active-verification rate	100.0%	50.0%

D CONFLICT-TRIGGERED FOLLOW-UP DIAGNOSTICS

This section examines adjudication replacement and conflict-triggered evidence acquisition with fixed first-round evidence. Table 2 gives all three point estimates; this section collects the independent data, implementation, paired statistics, and interpretive boundaries.

D.1 INDEPENDENT DATA AND SAMPLE SELECTION

Sample unit and scale of the supplementary diagnostic. The supplementary-module diagnostic uses $N = 200$ under each evidence condition. A sample is a literature question with a determinate reference claim. The same 200 questions are assembled under three evidence conditions, and each question–condition pair is evaluated with original evidence-priority synthesis, provenance-aware adjudication without follow-up, and conflict-triggered follow-up defined in Appendix D.3. The complete design therefore contains 600 question–condition combinations and 1,800 method outputs, rather than 1,800 independent questions. Table 2 reports all three configurations, and Appendix D.4 gives paired comparisons on the same per-question outputs without regenerating them. Fixed sample IDs preserve correspondence within and across conditions.

Candidate pool and source isolation. The supplementary data in `iclr_dataset.jsonl` contain 360 questions from 360 source families: 60 development questions and 300 candidate-confirmation questions. Construction excludes 1,591 source families appearing in the existing question banks and experimental records, yielding zero overlap with those sources. Source families collapse paper-version suffixes to prevent different versions of the same paper from crossing subsets. The complete 360-question pool contains 189 numerical, 161 comparative, and 10 conclusion claims; 200 candidate-confirmation questions are used in the reported diagnostic.

From source text to verifiable questions. Candidate claims are extracted from indexed body text after excluding reference sections. The builder first stores a verbatim source quotation and its character span, then generates a neutral question answerable from that quotation. Questions must identify the relevant entity, metric, and necessary experimental conditions without revealing the target value or conclusion direction, presupposing improvement, or omitting the comparison target. A reference claim represents what the source reports under the corresponding conditions; it is not treated as a universal truth across conditions.

Tampering and quality checks. Within the saved reference-claim span, construction changes one answer-relevant fact, such as a value, comparison, or conclusion. Associated wording within that span is adjusted when necessary, while the remaining source text is preserved. The false alternative must be incompatible with the authentic fact under the same conditions rather than another simultaneously valid bound. Separate calls check source support, question answerability, question neutrality, factual incompatibility, local fluency, condition preservation, and compatibility of benign variants. Rejected candidates retain their rejection reasons and are replaced from a presorted pool; screening and revision do not inspect the evaluated methods' answers.

Local construction records contain 822 body-text candidates. Processing the first 413 in fixed order produced 360 samples, with 35 API or structural errors and 18 semantic rejections in that prefix. A targeted text audit selected 104 samples for revision of answer leakage, mutual exclusivity of numerical ranges, consistency between values and wording, question–quotation correspondence, and preservation of unrelated text. The number 104 is the number selected for targeted review, not

the number confirmed to contain factual errors. Before-and-after quotations, replacements, reasons, and hashes are retained in the construction records.

Selection and review protocol. Dataset splitting and diagnostic-subset selection use seed 20260918. We fix 200 candidate-confirmation samples before method comparison, with one independent source family per question. Selection is independent of the outputs from the three configurations. Each question retains its source location, on-disk file SHA-256, clean and poisoned windows, and full replacement span. Data construction used the proxy service identifier `deepseek-flash`; generation and semantic checking were separate calls to the same configuration rather than human review or cross-model judging.

D.2 Fixed Evidence Assembly

Fixed evidence conditions. The local diagnostic initial context comprises one target snippet and four background snippets. The target is truncated to 4,000 characters. Background snippets are retrieved with BM25, excluding the target paper's source family, and each is truncated to 1,200 characters. One first external-evidence snippet is fixed, with contamination placed as specified in Table 14. For poisoned conditions, the external snippet is centered on the target claim and limited to 1,500 characters. The both-correct condition uses saved correct variants: identical original text, paraphrases, unit conversions, and compatible statements under different experimental conditions.

Table 14: Evidence assembly for supplementary diagnostics. P denotes the target tampering; C denotes correct evidence.

Condition	Initial target snippet	First external evidence
Poisoned initial / correct external (P/C)	Tampered text	Corresponding original support
Correct initial / poisoned external (C/P)	Clean text	Corresponding tampered snippet
Both correct (C/C)	Clean text	Correct, condition-compatible evidence variant

Contamination is applied to the initial or first external evidence supplied to the system; follow-up retrieval accesses the original offline literature index. Tampered snippets in the fixed-evidence diagnostic are not written into that index, so external contamination describes the first visible evidence configuration. Actual index injection uses a separate experimental pipeline. Configurations are paired by question and evidence condition, retaining their respective answers and judgments.

All three configurations share the same target snippet, background snippets, and first external evidence for a given question–condition pair; background content and order do not change across configurations. The P/C and C/P conditions use the same authentic–tampered pair and change only the location of the false fact. Models do not receive P/C or C/P labels, reference answers, tampering labels, or internal construction IDs. The provenance-aware branch receives only anonymous evidence indices and source-family groupings. The original synthesizer retains the role labels "initial context" and "deep evidence" because they are part of the original method's adjudication assumption, not extra truth labels introduced by the experiment.

The both-correct condition includes source-identical evidence, paraphrases, unit conversions, and compatible statements under different experimental conditions. Paraphrases and unit variants preserve the same fact; different-condition variants retain verbatim textual support and explicitly preserve the condition difference. Synthetic rewrites retain their supporting quotations and are not presented as verbatim source spans. This condition tests accuracy and unnecessary follow-up when no substantive conflict exists; it does not require all correct evidence to be worded identically.

D.3 Three Configurations and Shared Controls

Supplementary-module implementation. Supplementary diagnostics use the independent configuration in `iclr_protocol.json`. Table 15 gives the target-snippet and retrieval-chunk budgets for the 200-question evaluation. This configuration is recorded separately from the main-experiment settings in Table 7.

Table 15: Configuration for supplementary diagnostics and follow-up verification ($N = 200$).

Item	Local implementation configuration
Experimental scale	Same 200 questions × three evidence conditions × three configurations, yielding 1,800 method outputs
Initial context	One target snippet and four background snippets
Target / background length	At most 4,000 / 1,200 characters
First external evidence	One fixed snippet; at most 1,500 characters in poisoned conditions; saved variants when both are correct
Retriever	BM25
Follow-up trigger	Substantive-conflict flag from preliminary provenance-aware adjudication
Follow-up rounds / queries	At most one round / three queries
Candidate depth per query	Top-20
New evidence per round	At most six snippets
Retrieval chunk / overlap	1,500 / 200 characters
Within-round deduplication	Document ID, source family, and snippet content
Previously seen exclusion	Identical text hash, or same-document span overlap exceeding 0.5 of the shorter span
Previously seen papers	Unseen snippets from the same paper remain eligible
Final adjudication input	Initial, first external, and new evidence; the preliminary answer is not passed
Decoding	Temperature 0; maximum 3,000 output tokens per call
Generation model / endpoint	`deepseek-v4-flash` / official `https://api.deepseek.com/v1`
Judge model / budget	Separate call to the same official endpoint and model; temperature 0, maximum 1,200 output tokens
Model records	Configuration ID, service-returned model ID, and request records; a configuration name is not an immutable snapshot
Execution	Question-level parallelism, incremental result persistence, and failure retries; actual concurrency is recorded in the run manifest

Conflict assessment and follow-up evidence acquisition. The provenance-aware adjudicator aligns entities, metrics, units, versions, and experimental conditions, returning a preliminary answer, conflict flag, and dispute record. After a substantive-conflict trigger, the targeted branch generates three queries from the question, audited claim, and dispute record. Candidates are merged round-robin by within-query rank, deduplicated by document and source family within the round, and filtered to exclude seen text or highly overlapping snippets. Final adjudication uses the combined evidence and the same decision rules as preliminary adjudication. Without a trigger, the preliminary provenance-aware answer is reused. Source families organize versions and duplicates; they are not authenticity labels.

Compared configurations. Original evidence-priority synthesis uses the original AF-RAG evidence-priority synthesizer and its free-text output. All three configurations use the same generation model, temperature, and 3,000-token per-call output limit. Internal construction IDs are not passed to the model, and all evidence is rendered through the same evidence-display interface.

The no-follow-up and conflict-triggered configurations share the same preliminary adjudication and dispute record. Without a substantive conflict, follow-up directly reuses the no-follow-up answer. When triggered, final adjudication receives the combined evidence but not the preliminary answer, reducing direct anchoring. Original evidence-priority synthesis independently invokes AF-RAG's original synthesizer and does not depend on successful auditing or structured output from the provenance-aware branch. Auditing in that branch serves dispute-query construction and does not imply that original first-round retrieval has been executed.

D.4 Paired Results and Statistics

Supplementary diagnostic scoring. The judge separately assesses reference-claim adoption, target false-claim adoption, and benign-answer correctness. In poisoned conditions, adopting the reference claim without adopting the target false claim counts as joint success and contributes to JSR.

Table 16: Three configurations of the supplementary module. All configurations use the same question, initial evidence, and first external evidence.

Configuration	Adjudication	Follow-up retrieval	Reported in
Original evidence-priority synthesis	Evidence-priority synthesis	None	Table 2 and paired tests
Provenance adjudication (no follow-up)	Provenance-aware adjudication	None; returns the preliminary answer	Table 2 and paired tests
Conflict-triggered follow-up	Same provenance-aware adjudication as the no-follow-up configuration	At most one targeted round after a substantive-conflict trigger, followed by readjudication	Table 2 and paired tests

Both-correct conditions report accuracy from benign-answer correctness. Quoting, negating, or attributing a false claim to another source does not automatically count as adoption. The absence of a valid system answer counts as failure under the fixed denominator.

Independent scoring and statistical protocol ($N = 200$ per condition). This protocol applies to original evidence-priority synthesis, provenance-aware adjudication without follow-up retrieval, and conflict-triggered follow-up. It does not reuse the main evaluation's samples, confidence intervals, or p values. In poisoned conditions, T_i indicates adoption of the reference fact under the correct conditions and P_i indicates adoption of the designated false alternative. Joint success is $Y_i = T_i(1 - P_i)$, with $\text{JSR} = 100 \sum_i Y_i/200$. Here $P_i = 0$ means that the binary judge did not mark false-claim adoption; it is not the main evaluation's separate explicit-non-adoption label. The two JSR operationalizations and their results should therefore not be pooled or compared as if identically scored. Both indicators may be true: neither TRR nor $1 - \text{PAR}$ can substitute for joint correctness. Both-correct conditions use a separate benign-answer-correctness field and the same fixed denominator.

Answers are paired by question ID and evidence condition. Percentage-point differences use 5,000 paired bootstrap resamples by source family with seed 20260918. When each question has an independent source family, this is equivalent to question-level paired resampling, with a two-sided exact McNemar test. If multiple questions share a family, clustered intervals preserve within-family dependence. The nine predefined configuration contrasts across the three conditions form one Holm-corrected family.

Table 17 reports two mechanism comparisons per condition. Left/right wins count instances where only the left/right configuration succeeds. For follow-up minus provenance adjudication, these are corrections and regressions. With complete paired judgments, the net gain is $100(n_{\text{left}} - n_{\text{right}})/200$ points; it cannot replace the transition counts or establish an absence of regressions.

Table 17: Paired mechanism comparisons for the supplementary module ($N = 200$ per condition). Differences are left minus right in percentage points. Original denotes evidence-priority synthesis; provenance denotes adjudication without follow-up.

Initial / external	Comparison	Difference	Unadjusted 95% bootstrap CI	Left/right wins	Holm p
Poisoned / correct	Provenance − original	-13.5	$[-25.5, -1.0]$	66/93	0.233
	Follow-up − provenance	$+6.0$	$[+1.0, +11.5]$	21/9	0.233
Correct / poisoned	Provenance − original	$+19.0$	$[+13.5, +24.5]$	39/1	5.97×10^{-10}
	Follow-up − provenance	$+16.0$	$[+9.5, +22.5]$	41/9	3.93×10^{-5}
Both correct	Provenance − original	-2.0	$[-4.0, -0.5]$	0/4	0.500
	Follow-up − provenance	$+0.5$	$[0.0, +1.5]$	1/0	1.000

Intervals are unadjusted source-family paired-bootstrap percentile intervals. Two-sided exact McNemar p values receive Holm correction over all nine predefined contrasts. Table 17 shows six

mechanism comparisons; the three follow-up-versus-original total-effect contrasts enter that correction but their intervals and adjusted p values are not listed here. Their point estimates in Table 2 do not replace paired statistics. After correction, both provenance adjudication and follow-up retain significant gains under correct initial/poisoned external evidence; the other four mechanism comparisons are not significant. Unadjusted intervals and adjusted discrete tests may differ; significance claims follow the Holm-adjusted McNemar tests.

Failures, missingness, and judge reliability. API and structural-parsing failures remain in the fixed denominator of 200 and count as unsuccessful. Paired tests report missing pairs separately rather than treating a missing judgment as adoption of a false fact. Benign accuracy is evaluated against the reference fact and experimental conditions; an answer is not automatically correct merely because it avoids the poisoned content.

D.5 Interpretive Boundaries

Comparing provenance-aware adjudication with original synthesis replaces the entire adjudication configuration, including prompts, source-group presentation, and output structure. It is not a pure ablation of one provenance field. Comparing conflict-triggered follow-up with no-follow-up adjudication isolates targeted retrieval and readjudication as a combination; comparison with original synthesis measures the total extension effect. The three configurations do not separate targeted queries, neutral follow-up retrieval, and repeated adjudication, and do not establish superiority over equal-budget generic queries.

Because the diagnostic fixes first-round evidence acquisition, it measures adjudication and follow-up verification conditional on a supplied evidence configuration; it cannot estimate the probability that the original system initially finds corrective evidence. Correct sources remain in the follow-up retrieval corpus but are not supplied to the generator as annotated answers. Actual retrieval still depends on queries, ranking, chunking, and deduplication. Conclusions do not extend to unreachable correct sources, simultaneous contamination of first-round and follow-up retrieval, or compromise of the entire literature index.

E Core Mechanisms and Statistical Evidence

E.1 Paired Tests of Main Results

Table 18: Complete main security results under the only-poison cocoon ($N = 1,200$). Vanilla LLM is a closed-book reference; JSR requires truth recovery without adopting the poisoned claim.

Method	PAR ↓	TRR ↑	JSR ↑
Vanilla LLM (closed-book)	4.00%	30.25%	28.00%
Naive RAG	94.75%	25.58%	3.75%
GraphRAG	94.00%	31.08%	3.25%
RAGuard	94.75%	24.92%	3.83%
FLARE-style RAG	95.17%	26.08%	3.42%
Blind Retry RAG	90.00%	33.25%	7.67%
Retrieve-More RAG	84.67%	40.92%	10.33%
Query-Rewrite RAG	85.08%	43.08%	11.25%
Claim-Only RAG	85.33%	43.58%	10.17%
AF-RAG	**42.67%**	**67.50%**	**49.83%**

This section uses the same main-experiment records and scoring rules described in Appendix A.4.

Paired statistics. Method outputs are aligned by instance in the file. We perform 20,000 paired bootstrap resamples of per-instance binary metric differences using seed 20260919 and report percentile 95% intervals. McNemar tests use the two-sided exact binomial test; reported p values are not corrected for multiple comparisons. Let U denote the uncertain fraction. Then

$$\max(0, \mathrm{TRR} - \mathrm{PAR} - U) \le \mathrm{JSR} \le \min(\mathrm{TRR}, 1 - \mathrm{PAR} - U).$$

Table 19: Paired tests relative to Naive RAG ($N = 1,200$). Panels report PAR avoidance, TRR, and JSR. Intervals are for differences in percentage points.

Method	Rate	Change	95% CI	McNemar p
PAR avoidance				
AF-RAG	57.33%	+52.08pp	[+49.08, +55.08]	1.562×10^{-167}
Claim-Only RAG	14.67%	+9.42pp	[+7.67, +11.17]	4.791×10^{-27}
Query-Rewrite RAG	14.92%	+9.67pp	[+7.83, +11.58]	1.152×10^{-25}
Retrieve-More RAG	15.33%	+10.08pp	[+8.25, +12.00]	5.559×10^{-27}
TRR				
AF-RAG	67.50%	+41.92pp	[+39.00, +44.83]	2.568×10^{-130}
Claim-Only RAG	43.58%	+18.00pp	[+15.67, +20.33]	3.387×10^{-52}
Query-Rewrite RAG	43.08%	+17.50pp	[+15.25, +19.75]	1.890×10^{-53}
Retrieve-More RAG	40.92%	+15.33pp	[+13.17, +17.58]	3.435×10^{-42}
JSR				
AF-RAG	49.83%	+46.08pp	[+43.08, +49.00]	1.188×10^{-145}
Claim-Only RAG	10.17%	+6.42pp	[+4.92, +8.00]	6.606×10^{-17}
Query-Rewrite RAG	11.25%	+7.50pp	[+5.83, +9.17]	2.263×10^{-19}
Retrieve-More RAG	10.33%	+6.58pp	[+5.00, +8.17]	5.521×10^{-17}

E.2 Component Ablations

The ablation study is conducted on a fixed $N = 500$ subset drawn from the same evaluation pool as the main only-poison cocoon setting. The main robustness and significance results retain the full $N = 1200$ setting; the smaller ablation subset is used only to isolate component contributions under repeated API-bound evaluation runs.

Table 20 reports the contribution of each AF-RAG component. The full variant reaches 47.2% PAR and 61.2% TRR on the ablation subset. Removing the claim auditor produces a moderate degradation, removing the deep retriever collapses performance close to Naive RAG, and removing the evidence-prioritized synthesizer sharply reduces truth recovery despite retrieval being available.

Table 20: AF-RAG component ablation results (N=500).

Variant	PAR ↓	TRR ↑
Full AF-RAG	47.2%	61.2%
w/o claim auditor	61.4%	55.2%
w/o deep retriever	94.0%	10.0%
w/o synthesizer	84.4%	22.2%

E.3 Query-Derived Factual-Target Control

Query-derived factual-target control. This baseline derives a neutral factual target solely from the user query. The independent run uses a distinct model snapshot and sample configuration; its point estimates are compared only within Table 21 and are separate from the $N = 1200$ main evaluation.

Table 21: Independent query-derived factual-target control.

Method	PAR ↓	TRR ↑
Naive RAG	93.0%	11.5%
Claim-Only	82.5%	23.0%
Query-derived Factual Target	77.5%	15.0%
AF-RAG	**59.0%**	**60.5%**

E.4 Parametric Memory and Evidence-Source Controls

We test whether recovery can be explained by model memory, arbitrary added context, or target-document leakage on a fixed $N = 200$ subset.

Table 22: Parametric-memory and evidence-source controls ($N = 200$).

Metric	Closed-book	Poison-only	Random-deep	Target-removed	Full
PAR ↓	1.5%	95.5%	93.0%	89.5%	58.5%
TRR ↑	13.5%	18.0%	37.0%	32.0%	58.0%

Closed-book TRR is only 13.5%, and removing the target corrective document lowers TRR to 32.0%. The full-system gain therefore cannot be explained entirely by parametric memory or arbitrary additional context. The 37.0% and 32.0% TRR under Random-deep and Target-removed also show that parametric knowledge and non-target context remain non-negligible contributors.

F Applicability Boundaries and Robustness

F.1 Evidence-Placement Boundary

We directly test the local corrective-evidence assumption by changing evidence placement in a fixed $N = 200$ subset.

Table 23: Evidence-placement boundary test ($N = 200$); all values are percentages.

Initial Top-K	Deep pool	Method	PAR / truth accuracy
Poison	Corrective	AF-RAG	55.0 / 57.5
Clean	Poison	Naive RAG	0.5 / 97.0
Clean	Poison	AF-RAG	60.0 / 17.0
Poison	Poison	AF-RAG	87.5 / 15.5

When the initial context is clean but the deep pool is poisoned, AF-RAG reduces truth accuracy from the Naive RAG value of 97.0% to 17.0%. When both locations are poisoned, AF-RAG likewise fails to recover the truth. AF-RAG should therefore not be deployed as an unconditional always-on defense without source verification or evidence-placement gating.

F.2 RAGuard with Visible Contrastive Evidence

The only-poison cocoon setting evaluates defense behavior when authentic counter-evidence is absent from the initial context. This setting does not fully characterize defenses that require inter-document comparison. To avoid conflating this threat-model constraint with an implementation defect of RAGuard, Table 24 reports a separate fair setting in which each input context contains the poisoned document, the authentic source document, and benign distractors.

Table 24: Supplementary fair-setting results for isolation-based defense when poisoned and authentic contrastive documents are both visible in the input context (N=1200).

Architecture	PAR ↓	TRR ↑
Naive RAG	82.9%	24.1%
RAGuard	45.3%	34.8%
AF-RAG	54.1%	56.8%

Table 24 shows that RAGuard reduces PAR from 82.9% to 45.3% when authentic contrastive documents are present, confirming that isolation-based filtering can suppress poisoned adoption when the required comparison evidence is available. AF-RAG obtains a higher PAR of 54.1% in this fair setting because it does not physically remove the poisoned passage before synthesis; instead, the

synthesizer must adjudicate among the initial poisoned passage, the visible authentic passage, and deep-retrieved evidence. This makes poisoned-claim suppression less aggressive than RAGuard's deletion-based strategy, although AF-RAG still achieves higher answer-level truth recovery (56.8% TRR vs. 34.8%).

Table 25: Isolation diagnostics under the RAGuard fair setting.

RAGuard diagnostic	Value
Authentic source present	1200/1200 (100.0%)
RAGuard active	1200/1200 (100.0%)
Dropped poison doc index 0	889/1200 (74.1%)
Dropped authentic index 1	162/1200 (13.5%)

Table 25 complements Table 24 by showing which source leave-one-out isolation removes. This separates poisoned-document removal from answer-level truth recovery.

F.3 Tampering Consequences and Detectability

Table 26: Micro-tampering consequence controls ($N = 200$ per subset). The controls are separate from the $N = 1200$ main evaluation.

Subset	Metric	Naive RAG	AF-RAG
Query-consequential	PAR ↓	91.5%	59.0%
Query-consequential	TRR ↑	11.0%	58.5%
Query-non-consequential	Unnecessary conflict ↓	–	3.5%

The 3.5% unnecessary-conflict rate shows that AF-RAG does not indiscriminately trigger on every numerical or wording difference. This control tests consequence and trigger selectivity and does not replace the main evaluation.

The noisy-versus-stealth diagnostic evaluates how attack detectability changes robustness. On the noisy subset ($N = 50$), AF-RAG reduces PAR to 38.0% and reaches 72.0% TRR. On the stealth subset ($N = 50$), Naive RAG becomes substantially more vulnerable, reaching 94.0% PAR and only 8.0% TRR, while AF-RAG preserves 62.0% TRR. This contrast supports the claim that local self-consistency and the absence of visible conflict increase attack strength.

F.4 Cross-Domain Robustness

We apply the same construction and evaluation pipeline to 500 papers and 200 query–claim instances from each of arXiv Physics and the PubMed Central Open Access Subset.

Table 27: Cross-domain results on Physics and PMC Biomedical ($N = 200$ per domain).

Method	Physics PAR ↓	Physics TRR ↑	PMC PAR ↓	PMC TRR ↑
Naive RAG	75.0%	31.0%	85.0%	14.0%
GraphRAG	76.0%	52.0%	84.0%	34.5%
RAGuard	73.5%	29.0%	81.5%	15.5%
Blind Retry	68.0%	40.5%	76.0%	23.0%
Query-Rewrite	67.0%	45.5%	78.5%	26.0%
Claim-Only	64.0%	48.0%	74.0%	24.5%
AF-RAG	**42.5%**	**69.0%**	**54.5%**	**60.0%**

AF-RAG obtains the lowest PAR and highest TRR in both supplements. These datasets reuse the same controlled tampering and evidence-placement protocol, however, so the result tests trend transfer rather than open-world generalization.

F.5 Multi-Model Robustness

All configurations use the same corpus, retrieval settings, prompts, temperature 0, and $N = 200$ paired instances; only the listed LLM components change.

Table 28: Independent multi-model robustness evaluation ($N = 200$ paired instances).

Auditor/Generator	Synthesizer	Judge	Method	PAR ↓	TRR ↑
DeepSeek-V4-Flash	DeepSeek-V4-Flash	DeepSeek-V4-Flash	Naive RAG	94.5%	11.0%
DeepSeek-V4-Flash	DeepSeek-V4-Flash	DeepSeek-V4-Flash	AF-RAG	47.5%	59.0%
GPT-5.4-mini	GPT-5.4-mini	GPT-5.4-mini	Naive RAG	94.0%	12.0%
GPT-5.4-mini	GPT-5.4-mini	GPT-5.4-mini	AF-RAG	44.0%	58.5%
Qwen3.6-plus	Qwen3.6-plus	Kimi-K2.6	Naive RAG	93.0%	12.5%
Qwen3.6-plus	Qwen3.6-plus	Kimi-K2.6	AF-RAG	42.0%	59.5%

AF-RAG improves over Naive RAG under every listed generator and judge configuration, reducing the likelihood that the result is an artifact of a single model or same-model judge.

G Failure Analysis and Benign-Query False Alarms

G.1 Retrieval, Truth Recovery, and Joint Success

Table 29: Pipeline diagnostics under the only-poison cocoon setting.

Diagnostic	AF-RAG	Blind Retry	RAGuard
Audit / active	1200/1200 (100.0%)	–	0/1200 (0.0%)
Deep search	1200/1200 (100.0%)	–	–
Target hit	1149/1200 (95.75%)	933/1200 (77.75%)	–

Table 29 reports audit activation, deep-search triggering, and target reachability under the only-poison cocoon to distinguish active falsification from repeated retrieval. RAGuard isolation with visible contrastive evidence appears in Appendix F.2.

Poisoned-claim adoption and truth recovery are judged independently. Table 30 retains an uncertain column to distinguish explicit non-adoption from indeterminate judgments. Among 810 truth-recovering answers, 598 explicitly do not adopt the poisoned claim, 206 still adopt it, and six are uncertain. Joint success is therefore 598/1,200=49.83%.

Table 30: Contingency between AF-RAG truth recovery and poisoned-claim adoption. Percentages are relative to all 1,200 samples.

Truth recovered	Adopted	Not adopted	Uncertain	Row total
Yes	206 (17.17%)	598 (49.83%)	6 (0.50%)	810 (67.50%)
No	306 (25.50%)	77 (6.42%)	7 (0.58%)	390 (32.50%)
Column total	512 (42.67%)	675 (56.25%)	13 (1.08%)	1,200 (100.00%)

G.2 Stage-Specific Failures to Recover Truth

Table 31 follows the staged classification in `analyze_failure_modes.py`: audit, query-generation, and retrieval triggers are checked first, followed by target hits and answer labels. Target hits use `trr_metrics.afrag_trr`, a retrieval diagnostic; answer-level truth recovery uses `truth_recovery.afrag`. Failure percentages use the 390 samples without truth recovery as their denominator. There are 602 joint failures in total, additionally including the 206 poisoned-adoption and six uncertain answers in the first row of Table 30.

Table 31: Taxonomy of failures to recover truth ($N = 390$). All-sample percentages use $N = 1,200$ as the denominator.

Failure type / metric	Count	All samples	Failures
Audit not triggered	0	0.00%	0.00%
Query generation failed	0	0.00%	0.00%
Deep search not triggered	0	0.00%	0.00%
Truth not retrieved	27	2.25%	6.92%
Truth retrieved but answer still adopts poison	291	24.25%	74.62%
Truth retrieved but answer not recovered	72	6.00%	18.46%
Judge or trace error	0	0.00%	0.00%

Of the samples without truth recovery, 363/390 (93.08%) already hit the authentic target. This result identifies post-retrieval snippet selection, condition alignment, and answer synthesis as the primary directions for improvement.

G.3 Benign-Query False-Conflict Breakdown

A benign-query false-conflict breakdown attributes most AF-RAG alerts to numerical or metric-bearing claims (136/196, 69.4%), followed by broad factual claims (37/196, 18.9%) and comparative claims (15/196, 7.7%). Intent-guard misses account for 8/196 (4.1%), indicating over-sensitive verification of substantive claims rather than systematic triggering on summary or bibliographic intents.

These 196 false conflicts belong to the main evaluation and are kept separate from the independent FCR estimates below.

G.4 Complete Benign-Query Utility Results

Table 32 gives the complete results underlying Table 1. Accuracy changes are percentage points relative to Naive RAG. FCR and FRR are distinct signals and are not directly interchangeable.

Table 32: Complete benign-query utility results ($N = 1,200$). FCR denotes false conflict rate; FRR denotes RAGuard's false rejection rate.

Architecture	Accuracy ↑	Δ Accuracy (pp)	Signal
Naive RAG	63.7%	0.0	–
GraphRAG	67.8%	+4.1	–
Naive+Prompt	64.0%	+0.3	FCR 12.8%
Multi-Agent	68.8%	+5.1	FCR 3.0%
RAGDefender	61.1%	-2.6	–
ReliabilityRAG	62.4%	-1.3	–
RAGuard	2.5%	-61.2	FRR 96.6%
AF-RAG	**79.9%**	**+16.2**	FCR 16.3%

G.5 Independent Benign-Utility Replication

The following intervals come from an independent evaluation. They are not confidence intervals for Table 1 or Table 32 and are not pooled with the false-conflict breakdown in Appendix G.3.

Table 33: Benign-utility and false-positive intervals from an independent run, not interval estimates for Table 1.

Metric	Method / comparison	Result
Benign accuracy	Naive RAG	63.92% [61.16, 66.59]
Benign accuracy	AF-RAG	80.75% [78.42, 82.88]
Benign accuracy	AF-RAG vs. Naive RAG	+16.83 pp, 95% CI [+14.42, +19.33], $p = 7.35\times10^{-40}$
FCR	Naive+Prompt	12.42% [10.67, 14.40]
FCR	AF-RAG	15.75% [13.80, 17.92]
FCR	AF-RAG vs. Naive+Prompt	+3.33 pp, 95% CI [+1.00, +5.67], $p = 6.98\times10^{-3}$
FRR	RAGuard	97.08% [95.97, 97.90]