REVISIONSCOPE: INTENDED-TIME GROUNDING FOR VERSIONED SECURITY ADVISORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

A security citation can entail every generated claim and still prescribe the wrong mitigation: the cited advisory is authentic, but superseded at the time for which the answer must hold. A revision-grounded evaluation makes this failure measurable by pairing each question with an authoritative intended time and an answer-changing revision boundary, applying four interventions at the query, metadata, update, and cache interfaces, and scoring the cited revision directly with judge-free temporal citation validity (TCV). RevisionScope turns the same evaluation object into a scope-carrying inference invariant: one resolved scope, corpus epoch, stable revision chain, applicability state, and cross-revision conflict record govern retrieval and generation. On 2,144 public SecAdvisory and CVEFixes-DB questions, plain RAG attains 91.6 [0.9]% content support while reaching only 27.6% TCV and 73.8% attack success rate (ASR). RevisionScope raises TCV to 92.3% and lowers ASR to 9.4%, contracting the support–scope gap from 64.0 to 3.8 points; the strongest measured source-reliability comparator reaches 85.4 [1.5]% TCV and 14.8 [1.9]% ASR with higher latency and abstention. The operating point also preserves 95.5 [0.8]% temporal answer accuracy with 1.6% abstention on ordinary queries and transfers without threshold tuning to 640 Kubernetes questions at 93.1 [1.0]% TCV and 7.5 [1.2]% ASR. Across 4,288 questions from four versioned corpora, scope-carrying inference raises TCV from 28.4% to 93.2% and lowers ASR from 72.8% to 8.6% at 18.8% P50 latency overhead. Intended-time grounding therefore changes both system selection and the safety meaning of a supported citation.

1 INTRODUCTION

Figure 1 traces a three-revision advisory. A responder asks for the current mitigation after $v2.0$ becomes effective, but a competing date cue steers retrieval to $v1.0$. The assistant quotes authentic vendor text, cites it faithfully, and recommends the obsolete action—a failure conventional grounding still accepts. On 2,144 public security questions, plain RAG obtains 91.6 [0.9]% content support, yet only 27.6% of its citations match the intended answer time and temporal interventions make 73.8% of answers unsafe. We explain and control this 64.0-point support–scope gap.

The graph also separates requests that recency conflates. A legitimate historical question binds to an old revision; a present-scoped request containing quoted, subordinate, or injected history remains bound to its current task clause. We encode that authoritative scope as $t_q^\star$, fixed before intervention, and distinguish it from the system-resolved $\hat{t}_q$. For citation d with applicability interval $I(d)$, temporal integrity requires $t_q^\star \in I(d)$ and action agreement at $t_q^\star$. Revision metadata determines the first condition without an answer judge; the second records operational consequence.

Freshness mishandles historical questions by always preferring newer evidence. Temporal RAG usually identifies the requested date with the time consumed downstream; provenance and reliability establish source or support even though one trusted source can issue incompatible revisions. Our protocol instead follows one advisory through scope resolution, revision selection, citation, and action to expose where intended time is lost. The trace matters operationally because mitigations for incidents such as Log4Shell changed across advisory revisions (Apache Logging Services, 2026; Apache Software Foundation, 2021).

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.

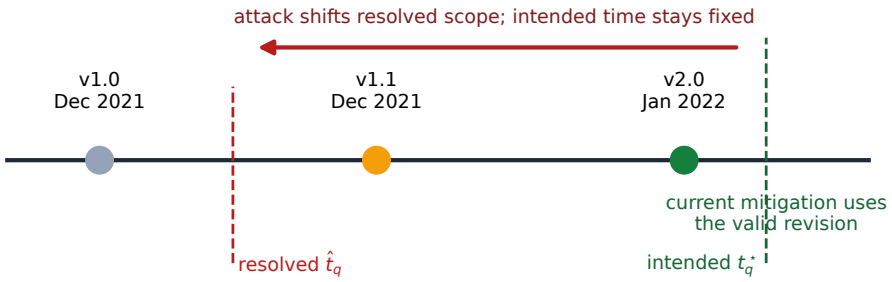

Figure 1: One advisory, three authentic revisions, and two time variables. A legitimate historical request moves both $t_q^\star$ and $\hat{t}_q$; scope steering moves only the system's $\hat{t}_q$.

Four capability-indexed interventions target the revision graph: query backdating inserts a non-authoritative historical cue; timestamp spoofing changes order metadata; update poisoning introduces an action-changing revision; and stale-cache replay returns an obsolete snapshot. Questions anchor to adjacent revisions with different recommendations, fixing acceptable revisions and actions before perturbation. TCV tests the cited revision; ASR tests whether mismatch becomes unsafe advice. Ordinary queries, separately authored prompts, adaptive attacks, and Kubernetes documents probe utility and transfer.

RevisionScope implements *scope-carrying inference*: one root-task scope remains immutable across retrieval, applicability verification, caching, and generation. Retrieval constructs a valid-at-scope frontier over stable document identities; the verifier preserves uncertain applicability; the cache binds results to revision state; and generation receives the same scope and explicit cross-revision conflict state. This invariant converts the evaluation trace into a system design aligned with the four intervention interfaces.

The public comparison tests this mechanism against source credibility, factual verification, corroboration, and added inference. RevisionScope reaches 92.3% TCV and 9.4% ASR, versus 85.4 [1.5]% and 14.8 [1.9]% for the strongest source-reliability comparator, with lower P50 (279 vs. 329 ms) and abstention (5.6% vs. 8.3%). It attains 95.5 [0.8]% accuracy with 1.6% abstention on 800 ordinary queries and, without threshold tuning, 93.1 [1.0]% TCV and 7.5 [1.2]% ASR on 640 Kubernetes questions. Across four corpora, it moves TCV from 28.4% to 93.2% and ASR from 72.8% to 8.6%; three downstream component removals raise ASR to 17.4%, 31.8%, and 46.2%.

The result is a revision-level grounding unit, a stress protocol separating historical intent from scope steering, and an ablated end-to-end invariant. For versioned evidence, support cannot certify an answer until its citation is checked against the time the answer governs.

2 Related Work

RAG systems optimize relevance, answer support, and self-critique over external memory (Guu et al., 2020; Lewis et al., 2020; Karpukhin et al., 2020; Izacard et al., 2022; Khattab & Zaharia, 2020; Asai et al., 2024). VersionRAG routes version-specific, version-listing, and change queries over a document-history graph (Huwiler et al., 2025); T-GRAG decomposes multi-time questions over timestamped knowledge graphs (Li et al., 2025); NuggetIndex associates atomic facts with validity intervals and lifecycle state (Zerhoudi et al., 2026); and ChronoQA covers explicit, implicit, and multi-document temporal questions (Chen et al., 2025). Their primary object is temporally accurate retrieval or answering. Our object is the full revision trace from an independently fixed intended time to the citation and operational action, including interventions that separate intended from resolved scope. Matched Per-Doc Latest and Time Filter rows instantiate the relevant revision-selection primitives under the same candidates and generator.

Temporal QA and continual-knowledge benchmarks evaluate answer drift or knowledge adaptation across time (Liska et al., 2022; Jang et al., 2022; Saxena et al., 2021; Shang et al., 2022; Tan

Table 1: What is observed when authentic evidence is served at the wrong time.

Line of work	Time cond.	Rev. ID	Temp. attack	Cited-rev. label	Action
Temporal QA / continual knowledge	✓				
Temporal IR / T-GRAG / ChronoQA	✓				
VersionRAG / NuggetIndex	✓	✓			
Evolution / cache attacks		✓	✓		✓
Provenance / source reliability					✓
Temporal integrity	✓	✓	✓	✓	✓

et al., 2023; Meem et al., 2024). Temporal IR and time-aware RAG condition ranking on temporal intent (Han et al., 2025; Qian et al., 2024). In these formulations the requested time normally becomes the system's temporal condition. Our protocol retains both variables: it freezes $t_q^\star$, perturbs the channel that produces $\hat{t}_q$, preserves source identity and topic, and checks the revision actually cited. The resulting measurement distinguishes a correct historical answer from a present-scoped answer steered toward the same old revision.

RAG poisoning and prompt-injection studies manipulate retrieved content or instructions (Greshake et al., 2023; Liu et al., 2023; Wang et al., 2025; Zou et al., 2025; Chaudhari et al., 2024; Carlini et al., 2024). KEPo is especially relevant because it forges plausible knowledge-evolution paths in GraphRAG (Chen et al., 2026); our update intervention instead follows a stable advisory identity and jointly labels revision applicability and the changed action. Recent cache-security work attacks semantic or prefix reuse through crafted collisions and shared serving state (Wu et al., 2026a;b), while GroundedCache validates cached answers against freshly retrieved evidence and source versions (Shah, 2026). Stale-cache replay complements these settings by asking whether a versioned retriever serves the revision required by a fixed intended time under an expected corpus epoch.

Provenance, credibility, source-reliability, and conflict-modeling methods score factual support, source trust, or cross-passage agreement (Sankararaman et al., 2024; Pan et al., 2024; Hwang et al., 2025; Zhang et al., 2025). We instantiate Provenance and RA-RAG with source metadata and their factual-support or reliability signals while withholding revision-validity labels. Recency, Per-Doc Latest, and Time Filter isolate increasingly informed version-selection primitives under the same candidate pool; Agentic-RAG and Multi-Source test added inference and corroboration. This comparison determines which gains require intended-time revision state rather than credibility, support, recency, or compute.

Table 1 locates the technical distinction in what is observed jointly: authoritative intent, resolved scope, cited revision, and action outcome under controlled perturbation. RevisionScope mirrors that object at inference time by preserving one scope and revision chain across stages. Table 4 shows the resulting evaluation reversal: content support remains high while cited-revision validity and answer safety collapse.

3 Threat Model

We study a versioned RAG assistant with ingestion, retrieval, generation, and caching stages. Each revision d has text x_d, a stable document key $\mathrm{id}(d)$, an ingestion-assigned ordered key $\mathrm{rev}(d)$, system-visible time metadata m_d, and a canonical applicability interval $I(d)$ used by the evaluation oracle. The inference stack receives either a source-declared interval or an estimate $\widetilde{I}(d)$ reconstructed from m_d, never the oracle label $I(d)$. A trusted control plane maintains stable identities, immutable

revision order, and monotone corpus epoch; source time fields and cache contents remain attack surfaces.

Task-scope contract. The request is represented as $q = (q_{\mathrm{root}}, q_{\mathrm{ctx}})$: q_{root} is the clause that asks the assistant to act, and q_{ctx} contains quoted, retrieved, or subordinate context. The intervention protocol preserves that functional boundary. A query adversary may alter q_{ctx} but cannot replace q_{root}; this models scope smuggling into an otherwise authenticated application request rather than arbitrary replacement of the user's task. The resolver sees the field boundary but never sees the gold revision or action.

Each instance has an authoritative intended time $t_q^\star$. For an explicit historical task, q_{root} names the historical date and $t_q^\star$ is that date; for a present-scoped task, $t_q^\star$ is the frozen session timestamp. Before intervention, the protocol fixes $t_q^\star$, acceptable revisions, reference action $y(q, t_q^\star)$, and the control plane's expected epoch $e_q^\star$. The system independently resolves $\hat{t}_q = R(q_{\mathrm{root}}, q_{\mathrm{ctx}})$ and copies $e_q = e_q^\star$ into its scope record; a cache entry carries its own tag e_c. A clean historical request has historical $t_q^\star = \hat{t}_q$. Backdating keeps a present-scoped root and $t_q^\star$ fixed while adding a competing cue to q_{ctx}.

Capability-indexed interventions. A *query adversary* supplies temporal language in q_{ctx} that can shift $\hat{t}_q$ away from $t_q^\star$. A *metadata adversary* alters m_d, and hence the system's estimated $\widetilde{I}(d)$, while x_d, $\mathrm{id}(d)$, and $\mathrm{rev}(d)$ remain fixed. A *corpus adversary* publishes an action-changing revision through an accepted source. A *serving adversary* substitutes a stale candidate payload, either with its old $e_c < e_q$ tag or beneath a surviving cache key; it cannot change the request's trusted e_q. These capabilities induce query backdating, timestamp spoofing, update poisoning, and stale-cache replay at four distinct interfaces.

Protected property. The evaluation oracle—$t_q^\star$, canonical $I(d)$, acceptable revisions, and reference action—remains outside inference. The root/context boundary, $\mathrm{id}(d)$, $\mathrm{rev}(d)$, e_q, and frozen model parameters are trusted; exactly one of query context, m_d, accepted-update semantics, or cached payload is varied per intervention. For the nonempty set $C(a)$ cited by answer a, temporal integrity is

$$\mathsf{TI}(q, a, C; t_q^\star) = 1 \quad \Longleftrightarrow \quad C(a) \neq \varnothing \ \wedge \ \big(\forall d \in C(a),\ t_q^\star \in I(d)\big) \ \wedge \ a \sim y(q, t_q^\star),$$

where $\sim$ denotes agreement on answer-changing content. Citation applicability is computed directly from stable revision identifiers and intervals; action agreement records the operational consequence. The two conditions separate wrong-time evidence from a malicious latest revision: poisoning can pass citation applicability while failing action agreement.

The taxonomy indexes capability rather than attacker identity: $\mathcal{A}_Q$ targets scope resolution, $\mathcal{A}_M$ applicability evidence, $\mathcal{A}_C$ accepted-update semantics, and $\mathcal{A}_S$ snapshot selection. Appendix Table 22 connects each capability to its protected object, trusted boundary, enforcement point, comparator, and measured residual.

4 REVISIONSCOPE: SCOPE-CARRYING RAG

RevisionScope enforces one invariant across the RAG pipeline: the scope resolved once from the root task governs candidate construction, cache admission, evidence applicability, and generated actions. A rule-based resolver emits $\sigma_q = (\hat{t}_q, m_q, \gamma_q, e_q)$: resolved time, explicit-historical or present-scoped mode, root-scope status $\gamma_q \in \{\mathrm{resolved}, \mathrm{ambiguous}\}$, and the control-plane epoch copied as $e_q = e_q^\star$. Retrieval, verification, and generation consume that record without reinterpretation:

$$\sigma_{\mathrm{retrieve}} = \sigma_{\mathrm{verify}} = \sigma_{\mathrm{generate}} = \sigma_q.$$

The evaluation oracle $t_q^\star$, acceptable revision set, and reference action are used only for scoring. The pipeline receives the request's root/context boundary and control-plane epoch, not a gold revision label. Downstream stages can attach uncertainty or conflict state, but they cannot silently change $\hat{t}_q$ or e_q; stable document and revision identities remain attached to every passage. Temporal state is therefore shared across interfaces instead of being re-inferred independently.

The resolver normalizes one root-clause date when present; with no root date it binds $\hat{t}_q$ to the frozen session timestamp, while incompatible root dates or an inseparable task clause set $\gamma_q = \mathrm{ambiguous}$

and force clarification. Temporal expressions in q_{ctx} remain evidence context, not scope overrides. Backdating modifies only that channel; spoofing changes source time metadata; poisoning adds one accepted action-changing revision; replay substitutes a cached payload. The other channels, topic, and $t_q^\star$ remain fixed.

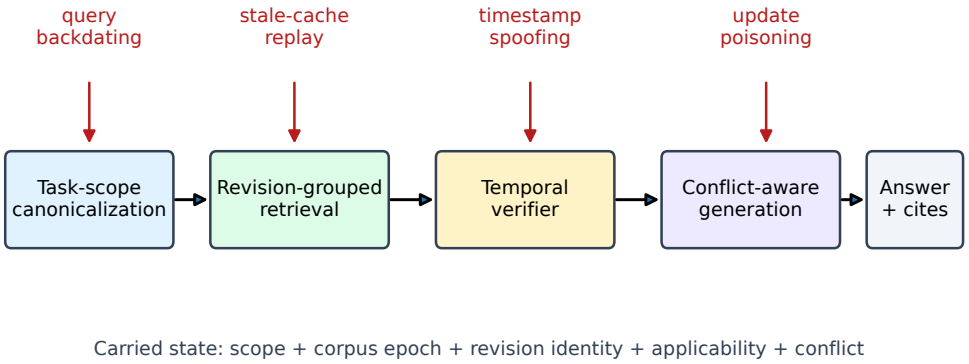

Figure 2: RevisionScope propagates one resolved scope and corpus epoch with stable revision identity, applicability, and conflict state.

Figure 2 exposes the information flow. Reuse first requires $e_c = e_q$; a mismatch is invalidated, while every surviving payload is regrouped by trusted $\mathrm{id}(d)$ and $\mathrm{rev}(d)$ before use. Retrieval constructs the valid-at-$\hat{t}_q$ frontier before semantic top-k selection. With source-declared intervals it tests $\hat{t}_q \in \widetilde{I}(d)$; with timestamp-only histories it estimates $\widetilde{I}(d)$ from m_d and carries uncertainty forward. The verifier scores (q, p_i, σ_q), returning applicability state plus pairwise supersession, effective-date, and action-conflict flags. The generator receives admitted passages, σ_q, and those structured fields.

Table 2: Scope-carrying inference operations.

Step	Operation
1	Resolve q_{root} once; emit immutable $(\hat{t}_q, m_q, \gamma_q, e_q)$.
2	Reject stale cache epochs; group candidates by identity and order revision chains.
3	Construct the frontier applicable to $\hat{t}_q$, retaining interval uncertainty.
4	Triage each (q, p_i, σ_q) with τ_{lo} and τ_{hi}.
5	Generate with the same scope and conflict state; clarify, compare, answer, or abstain.

The applicability verifier is a DeBERTa encoder (He et al., 2021; 2023) trained on 12,400 passage–time–label triples. Stable document identities are disjoint across 9,920 training, 1,240 calibration, and 1,240 triples. We use AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-5}, batch size 32, three epochs, and one NVIDIA A100-80GB GPU. After weights are frozen, calibration selects $\tau_{\text{hi}} = 0.65$ for acceptance and $\tau_{\text{lo}} = 0.35$ for rejection; intermediate scores remain uncertain. This model supplies applicability state, whereas the separate answer-level evaluator is calibrated on 400 human-labeled examples and never defines TCV. Section 5 tests the unchanged applicability thresholds on a different Kubernetes document family.

Conflict-aware generation follows an ordered policy. Ambiguous root scope requests clarification before action; explicit comparisons receive both revisions with intervals; one valid, conflict-free frontier supports an answer; unresolved present-scoped conflict produces abstention. For poisoning, the separately reported action-validation extension tests generated imperative steps against admitted evidence with a frozen DeBERTa model. A failed check triggers one constrained regeneration from that evidence, replacing the action and citations; failure to produce a validated action then abstains.

Tables report the final returned output, so correction can change both citation validity and action safety, whereas abstention scores zero on TCV. This extension targets partial-truth revisions that alter one operational instruction.

The ablation follows the same information path: revision-conditioned retrieval controls exposure, applicability verification controls passage, and conflict-aware generation controls whether surviving disagreement becomes an action. Appendix Table 15 measures these stages separately, while Table 21 tests malicious latest revisions that are temporally applicable but actionally unsafe. Together they distinguish revision-selection failures from unsafe semantics inside a valid revision.

5 EVALUATION

The experiments follow the causal path in Figure 2. *Q1* measures whether content support misclassifies authentic but wrong-time evidence. *Q2* tests whether freshness, version filtering, factual verification, source reliability, added reasoning, or individual RevisionScope components close that gap. *Q3* tests the resulting operating point on ordinary queries, independently authored prompts, adaptive attacks, and a different document family. Public SecAdvisory and CVEFixes-DB results carry the primary claim; the four-corpus analysis tests breadth.

The four-corpus evaluation contains 20,131 documents, 48,044 revisions, and 4,288 time-scoped questions. SecAdvisory (6,842 documents, 18,426 revisions) aggregates public vendor advisories, and CVEFixes-DB (Bhandari et al., 2021) links 5,284 public vulnerability records to 12,648 fixing revisions. PolicyDocs contributes 4,218 documents and 9,847 revisions; IncidentRunbooks contributes 3,787 documents and 7,123 revisions. Every corpus supplies 1,072 frozen questions, making the two-public-corpus aggregate an equal-weight 2,144-query evaluation.

Revision histories are normalized by stable document identity and ordered effective dates; consecutive effective dates delimit $I(d)$ when an explicit end date is absent. Construction begins from an adjacent pair whose operational recommendation changes. The question is retained only when one intended interval determines the requested action; $t_q^\star$, every revision whose interval contains it, and the reference action are then frozen before attacks are instantiated. Consequently, the TCV label is a deterministic lookup over the frozen intended time and normalized revision chain, independent of the learned answer judge. The reference action is used separately for TAA and ASR. Clean historical questions receive historical $t_q^\star$, whereas backdating preserves a present-scoped root request and changes only contextual temporal language.

Table 3: Intervention contract. Each row changes one channel after labels are frozen.

Intervention	Changed input	Held fixed
Backdate	Temporal cue in q_{ctx}	Root task, topic, corpus, metadata, cache
Spoof	Consumed publication/update order fields	Query, revision text/identity, cache
Poison	Action-changing text in one accepted update	Query, order, source identity, cache
Replay	Retrieval output from an earlier corpus epoch	Query, current revision chain, metadata

Strength increases adversarial leverage inside the named channel while preserving Table 3: Light changes one low-salience signal, Medium changes the directly consumed signal or record, and Strong coordinates the available signals within that channel. No strength grants a second channel or replacement of the root task. Medium is the headline regime. The challenge set estimates exploitability conditional on an answer-changing revision; the 800-query stratified set in Table 6 measures ordinary present-scoped behavior away from those boundaries.

The primary stack uses Llama-2-7B-Chat (Touvron et al., 2023), Contriever (Izacard et al., 2022), and FAISS (Johnson et al., 2021). Generator transfer covers Llama-3-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Phi-3-mini-4k-instruct (Abdin et al., 2024); retriever transfer covers BM25 (Robertson & Zaragoza, 2009), DPR (Karpukhin et al., 2020), and ColBERT-v2 (Santhanam et al., 2022). Experiments use PyTorch 2.1.2 and Transformers 4.36.2 on one NVIDIA A100-80GB GPU with CUDA 12.4 and driver 550.54.14. Results average seeds

42, 123, and 456. Each of 10,000 percentile-bootstrap replicates resamples frozen questions and carries their repeated-seed outputs together, preserving the pairing among methods (Efron, 1979). Brackets and the appendix's mean ± notation report 95% half-widths where shown. Latency uses five 500-query runs after 100 warm-up requests and reports end-to-end P50 and P95.

Applicability-verifier training, calibration, and test triples are partitioned by stable document identity. Attacked, ordinary, and held-out evaluations use disjoint question sets; held-out prompts are separately authored; and Kubernetes supplies a different document family evaluated with frozen SecAdvisory/CVEFixes thresholds. These splits test verifier generalization by identity, question-construction transfer in domain, and threshold transfer across document families.

For N queries, let s_i indicate that cited text supports the generated claim when revision time is hidden, v_i indicate that every citation in the final returned answer belongs to the valid set at $t^\star_{q_i}$, c_i indicate agreement with $y(q_i, t^\star_{q_i})$, u_i indicate an unsafe answer caused by the temporal intervention, and r_i indicate abstention. Missing citations and abstentions set $s_i = v_i = c_i = 0$; abstention sets $u_i = 0$. Action-validation rows are rescored after their single correction-or-abstention pass. We report

$$\mathrm{CS} = 100N^{-1}\sum_i s_i, \quad \mathrm{TCV} = 100N^{-1}\sum_i v_i,$$
$$\mathrm{TAA} = 100N^{-1}\sum_i c_i, \qquad \mathrm{AR} = 100N^{-1}\sum_i r_i.$$

and, on attacked queries, $\mathrm{ASR} = 100N^{-1}_{\mathrm{atk}}\sum_i u_i$. Content support (CS) reproduces the passage–claim entailment view; the *support–scope gap* is $\mathrm{SSG} = \mathrm{CS} - \mathrm{TCV}$. TCV is computed from nonempty cited revision identifiers and validity intervals without a learned judge. Because abstention sets $u_i = 0$, every attacked comparison reports AR beside ASR, making the safety–coverage operating point explicit. On ordinary queries, Correct is the raw count with $c_i = 1$, CiteP is citation precision over returned citations, and Abstain is the raw count with $r_i = 1$. A frozen DeBERTa-v3-large judge scores CS, TAA, and ASR; its threshold is fixed on a separate 400-example human calibration set before system comparison. Double annotation on that set reaches 94.2% agreement and $\kappa = 0.87$ (Cohen, 1960), with disagreements adjudicated. Appendix Table 17 audits answer-level ranking; all headline citation-validity claims remain judge-free.

The baseline family compares competing explanations. Recency reranks by update time; Per-Doc Latest retains one newest revision per stable identity; Time Filter applies $\hat{t}_q$ and a fixed look-back window. Agentic-RAG adds retrieval and reasoning calls, and Multi-Source requires corroboration. The Provenance adaptation applies one post-retrieval factual-support and citation-entailment pass (Sankararaman et al., 2024); the RA-RAG adaptation estimates reliability from public source metadata and answer-support signals before reranking (Hwang et al., 2025). Neither receives $t^\star_q$, applicability intervals, or revision-validity labels. Appendix A.3 fixes their information access, prompts, calls, context budgets, abstention behavior, and latency accounting.

5.1 Q1: Content support versus intended-time validity

Table 4: Public-corpus grounding disagreement under Medium-strength attacks on SecAdvisory + CVEFixes-DB ($n = 2,144$). CS hides revision time; SSG is CS–TCV; bracketed values report 95% bootstrap half-widths.

Method	CS↑	TCV↑	SSG↓	ASR↓
Plain RAG	91.6 [0.9]	27.6	64.0	73.8
RevisionScope	96.1 [0.6]	92.3	3.8	9.4

Table 4 produces a model-selection reversal. Content support classifies Plain RAG as 91.6% grounded, yet revision identity shows that 72.4% of its citations are invalid for the intended time. RevisionScope raises TCV by 64.7 points, contracts SSG by 60.2 points, and lowers ASR by 64.4 points under matched conditions. The joint movement of revision validity and unsafe outcomes makes the gap decision-relevant. Table 5 tests competing explanations; Appendix C supplies the blinded answer audit, separately authored validation, public poisoning, and per-attack counts.

5.2 Q2: Which interfaces close the gap?

Table 5: Public comparison under Medium-strength attacks ($n = 2,144$). Rows share queries, attacks, candidate pool, and context budget. Agentic-RAG uses its specified GPT-4o-mini controller; other rows use Llama-2-7B-Chat. Bracketed values report 95% bootstrap half-widths.

Method	TCV↑	TAA↑	ASR↓	AR	P50 (ms)
Plain RAG	27.6	33.6 [1.5]	73.8	0.0	235
Recency	44.9 [1.9]	50.7 [1.8]	53.1 [2.3]	0.0	241
Per-Doc Latest	53.6 [1.7]	58.9 [1.6]	44.7 [2.0]	0.0	246
Time Filter	69.1 [1.5]	74.4 [1.4]	32.6 [1.8]	1.7	257
Agentic-RAG	75.8 [2.3]	79.5 [2.1]	22.1 [2.6]	2.9	884
Provenance	78.9 [1.7]	80.6 [1.9]	20.8 [2.3]	6.4	304
Multi-Source	82.7 [1.6]	83.8 [1.8]	17.6 [2.2]	13.4	318
RA-RAG	85.4 [1.5]	86.2 [1.6]	14.8 [1.9]	8.3	329
RevisionScope	92.3	89.8 [1.0]	9.4	5.6	279

Table 5 compares intended-time state against matched credibility, factual-verification, corroboration, and extra-retrieval alternatives. RA-RAG is strongest at 85.4 [1.5]% TCV and 14.8 [1.9]% ASR; RevisionScope improves these by 6.9 and 5.4 points with 50 ms lower P50 and 2.7 points less abstention. Three downstream ablations test the information path: removing time-conditioned retrieval, applicability verification, or conflict-aware generation raises pooled ASR from 8.6% to 46.2%, 31.8%, or 17.4%, respectively (Table 15).

5.3 Q3: Utility, transfer, and adaptive pressure

Table 6: Benign utility on 800 ordinary present-scoped public queries sampled away from answer-changing boundaries. Counts use the frozen evaluation set; brackets denote 95% bootstrap half-widths.

Method	Correct	Abstain	TCV	TAA	CiteP
Plain RAG	736	0	91.1 [1.2]	92.0 [1.1]	92.6 [1.0]
Time Filter	744	9	94.3 [0.9]	93.0 [1.0]	94.8 [0.9]
RA-RAG	756	16	95.7 [0.8]	94.5 [0.9]	96.0 [0.8]
Multi-Source	750	35	96.1 [0.8]	93.8 [1.0]	96.4 [0.7]
RevisionScope	764	13	97.2 [0.6]	95.5 [0.8]	97.5 [0.6]

Method	AR	P50	P95	Latency units: ms
Plain RAG	0.0	228	508	
Time Filter	1.1	249	546	
RA-RAG	2.0	317	683	
Multi-Source	4.4	306	662	
RevisionScope	1.6	271	594	

The attacked gains do not trade away ordinary answer quality. Table 6 shows that RevisionScope attains the highest TAA and citation precision, improves TAA by 3.5 points over Plain RAG, and abstains on 13 of 800 requests. Plain RAG retains a 43 ms P50 advantage, while the guarded stack stays below 300 ms and avoids the high benign abstention of Multi-Source.

The transfer set freezes 160 questions per attack class and bootstraps paired repeated-seed outputs by question. Zero-shot evaluation reuses the SecAdvisory/CVEFixes verifier and thresholds; the last row tunes only those thresholds on a disjoint 64-example set. Without calibration, RevisionScope reduces unsafe answers from 342 to 48. Calibration changes TCV by 1.3 and ASR by 1.4 points, making zero-shot performance the decisive transfer result.

On the pooled four-corpus analysis ($n = 4,288$), RevisionScope reaches 93.2% TCV, 90.6% TAA, 8.6% ASR, 5.4% abstention, and 278 ms P50 (Appendix Table 8), beyond recency, latest-revision selection, filtering, agentic re-querying, and corroboration.

Table 7: Cross-family transfer on 640 Medium-strength attacked questions from public versioned Kubernetes release and operations documentation. Valid and unsafe are raw counts; brackets are 95% bootstrap half-widths.

Method	Valid	Unsafe	TCV	TAA	ASR	AR
Plain RAG	267	342	41.7 [2.2]	45.8 [2.4]	53.4 [2.7]	0.0
Time Filter	461	161	72.0 [1.8]	75.6 [2.0]	25.2 [2.2]	1.7
Provenance	503	125	78.6 [1.7]	80.6 [1.8]	19.5 [2.0]	5.8
RA-RAG	537	91	83.9 [1.5]	85.6 [1.6]	14.2 [1.8]	8.4
Multi-Source	543	88	84.8 [1.5]	85.1 [1.7]	13.8 [1.8]	12.2
RevisionScope (zero-shot)	596	48	93.1 [1.0]	90.8 [1.3]	7.5 [1.2]	5.0
RevisionScope (+64 cal.)	604	39	94.4 [0.9]	91.9 [1.2]	6.1 [1.0]	5.3

The ordering holds in every public attack channel: RevisionScope records 7.8/10.6/12.7/6.5% Backdate/Spoof/Poison/Replay ASR versus RA-RAG's 12.1/16.0/18.9/12.2% (Table 18). Poisoning is hardest because a latest revision can be applicable yet action-changing.

On 512 independently written public questions, RevisionScope reaches 94.8 [0.9]% TCV and 6.1 [1.0]% ASR, versus 41.8 [2.1]% and 46.9 [2.8]% for Plain RAG (Table 19). Under four disclosed-logic attacks, action validation lowers ASR from 24.0–33.7% to 16.3–22.9%, including 17.4 points on partial-truth edits (Table 21). These tests separate routing transfer from semantic validation of a malicious current revision.

6 Discussion

The 64.0-point support–scope gap changes grounding for evolving evidence: plain RAG cites entailing text, yet only 27.6% of public attacked cases cite the intended-time revision. Judge-free TCV therefore changes model selection by detecting when a valid-looking citation supports unsafe advice from the wrong scope.

Ablations locate the gain along the carried state: revision-conditioned retrieval controls exposure, applicability triage preserves temporal uncertainty, and conflict-aware generation blocks unresolved instruction. Unchanged-threshold Kubernetes transfer confirms that the interface contract crosses document families. A malicious current revision can still pass TCV; step-level entailment lowers public poisoning ASR from 18.4% to 11.6% and adaptive partial-truth ASR from 33.7% to 16.3%.

The evidence spans answer-changing boundaries and ordinary traffic in public security, Kubernetes, policy, and incident-response corpora; deployment prevalence depends on update cadence and request mix. The contract trusts the root/context boundary, stable revision identity, expected corpus epoch, and intended-time labels. TAA and ASR inherit audited-judge sensitivity, whereas central TCV is judge-free. Disclosed-logic attacks leave 16.3–22.9% ASR after action validation, prioritizing malicious current revisions and conflict suppression for stronger semantic defenses.

7 Conclusion

Versioned evidence makes citation grounding a temporal decision: which authentic revision governs the requested action? On 2,144 public security questions, that question exposes a 64.0-point gap behind 91.6 [0.9]% content support. The protocol measures it directly; RevisionScope carries scope, corpus epoch, revision chain, applicability, and conflict through generation to reach 92.3% TCV and 9.4% ASR, 95.5 [0.8]% ordinary-query TAA, and 93.1 [1.0]% zero-shot Kubernetes TCV. For evolving advisories, citation acceptance requires both textual support and authorization by intended time.

AI USE STATEMENT

Generative AI tools assisted in drafting and revising manuscript text, including the system name, title, abstract, and introduction, and in formatting and bibliographic consistency checks. The authors reviewed the AI-assisted manuscript preparation and take responsibility for the final content.

ETHICS STATEMENT

The attack taxonomy is dual use; its purpose is to identify and contain authentic-but-superseded evidence in retrieval-grounded assistants. All attacks were executed on controlled corpora and sandboxed stacks, and operational red-teaming should remain within explicitly authorized environments. Restricted enterprise documents and identifying metadata are excluded from public materials.

For that 400-example calibration exercise, annotators worked from a fixed rubric over static model outputs and revision-grounded evidence: they marked intended-time citation validity, agreement with the scoped reference action, and whether an intervention caused unsafe advice. The last label covers both advice supported by a wrong-time revision and, for poisoning, an action-invalid latest revision that can remain TCV-valid. Annotation was blinded to system identity during adjudication, disagreements were escalated to a third project member after independent labeling, and the labels calibrated only the answer-level judge rather than defining TCV or training RevisionScope. The exercise used internal annotators on non-personal, pre-existing corpus material rather than paid crowdworkers or live users; no separate compensation was provided, and no formal IRB or equivalent approval was sought.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the protected object, attacks, model and hardware configuration, metrics, seeds, uncertainty convention, and latency protocol. Appendix A.3 records exact baseline budgets and Appendix C provides raw-count, held-out, judge-audit, public-poisoning, and adaptive-attack analyses. The public evaluation boundary comprises Tables 4, 5, 6, 7, 11, 17, 18, 19, 20, and 21, together with the public rows of Table 9; Table 22 links each claim to these measurements. Pooled PolicyDocs and IncidentRunbooks results are reported separately.

REFERENCES

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Apache Logging Services. Security, 2026. URL https://logging.apache.org/log4j/2.x/security.html. Accessed: 2026-08-30.

Apache Software Foundation. CVE-2021-44228, 2021. URL https://www.cve.org/CVERecord?id=CVE-2021-44228. CVE Numbering Authority record; accessed 2026-08-30.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Guru Bhandari, Amara Naseer, and Leon Moonen. Cvefixes: Automated collection of vulnerabilities and their fixes from open-source software. In *Proceedings of the 17th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE)*, pp. 30–39. ACM, 2021. doi: 10.1145/3475960.3475985. URL https://doi.org/10.1145/3475960.3475985.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy*, pp. 407–425. IEEE, 2024. doi: 10.1109/SP54263.2024.00179. URL https://doi.org/10.1109/SP54263.2024.00179.

Harsh Chaudhari, Giorgio Severi, John Abascal, Anshuman Suri, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General backdoor attacks on retrieval augmented language generation, 2024. URL https://arxiv.org/abs/2405.20485.

Qizhi Chen, Chao Qi, Yihong Huang, Muquan Li, Rongzheng Wang, Dongyang Zhang, Ke Qin, and Shuang Liang. KEPo: Knowledge evolution poison on graph-based retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026*, pp. 2308–2319. ACM, 2026. doi: 10.1145/3774904.3792547. URL https://doi.org/10.1145/3774904.3792547.

Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. A question answering dataset for temporal-sensitive retrieval-augmented generation. *Scientific Data*, 12:1855, 2025. doi: 10.1038/s41597-025-06098-y. URL https://doi.org/10.1038/s41597-025-06098-y.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783v3.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, pp. 79–90. ACM, 2023. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

SeungYoon Han, Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, Huije Lee, and Jong C. Park. Temporal information retrieval via time-specifier model merging, 2025. URL https://arxiv.org/abs/2507.06782.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Daniel Huwiler, Kurt Stockinger, and Jonathan Fürst. VersionRAG: Version-aware retrieval-augmented generation for evolving documents, 2025. URL https://arxiv.org/abs/2510.08109.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34279–34303. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1738. URL https://aclanthology.org/2025.emnlp-main.1738/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.03215.

Albert Q. Jiang et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

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 (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

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

Dong Li, Yichen Niu, Ying Ai, Xiang Zou, Biqing Qi, and Jianxing Liu. T-GRAG: A dynamic GraphRAG framework for resolving temporal conflicts and redundancy in knowledge retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11880–11889. ACM, 2025. doi: 10.1145/3746027.3755628. URL https://doi.org/10.1145/3746027.3755628.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, Leo Yu Zhang, and Yang Liu. Prompt injection attack against LLM-integrated applications, 2023. URL https://arxiv.org/abs/2306.05499.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jannat Meem, Muhammad Rashid, Yue Dong, and Vagelis Hristidis. PAT-questions: A self-updating benchmark for present-anchored temporal question-answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13129–13148. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.777. URL https://aclanthology.org/2024.findings-acl.777/.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. Not all contexts are equal: Teaching LLMs credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19844–19863. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1109. URL https://aclanthology.org/2024.emnlp-main.1109/.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. TimeR^4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6942–6952. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.394. URL https://aclanthology.org/2024.emnlp-main.394/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. Provenance: A light-weight fact-checker for retrieval augmented LLM generation output. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1305–1313. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.97. URL https://aclanthology.org/2024.emnlp-industry.97/.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6663–6676. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.520. URL https://aclanthology.org/2021.acl-long.520/.

Syed Huma Shah. Grounded cache routing for retrieval-augmented generation: When is it safe to reuse an answer?, 2026. URL https://arxiv.org/abs/2605.27494.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8017–8026. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.552. URL https://aclanthology.org/2022.acl-long.552/.

Wang-Chiew Tan, Jane Dwivedi-Yu, Yuliang Li, Lambert Mathias, Marzieh Saeidi, Jing Nathan Yan, and Alon Y. Halevy. Timelineqa: A benchmark for question answering over timelines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 77–91. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.6. URL https://aclanthology.org/2023.findings-acl.6/.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Xilong Wang, John Bloch, Zedian Shao, Yuepeng Hu, Shuyan Zhou, and Neil Zhenqiang Gong. Webinject: Prompt injection attack to web agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2010–2030. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.104. URL https://aclanthology.org/2025.emnlp-main.104/.

Guanlong Wu, Taojie Wang, Yao Zhang, Zheng Zhang, Jianyu Niu, Ye Wu, and Yinqian Zhang. When cache poisoning meets LLM systems: Semantic cache poisoning and its countermeasures.

In *Proceedings of the 33rd Network and Distributed System Security Symposium*. Internet Society, 2026a. doi: 10.14722/ndss.2026.240200. URL https://doi.org/10.14722/ndss.2026.240200.

XiangFan Wu, Lingyun Ying, Guoqiang Chen, Yacong Gu, and Haipeng Qu. Cache me, catch you: Cache related security threats in LLM serving frameworks. In *Proceedings of the 33rd Network and Distributed System Security Symposium*. Internet Society, 2026b. doi: 10.14722/ndss.2026.242812. URL https://doi.org/10.14722/ndss.2026.242812.

Saber Zerhoudi, Michael Granitzer, and Jelena Mitrović. NuggetIndex: Governed atomic retrieval for maintainable RAG. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2286–2296. ACM, 2026. doi: 10.1145/3805712.3809687. URL https://doi.org/10.1145/3805712.3809687.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062/.

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*, pp. 3827–3844. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A Additional Experimental Details

A.1 Pooled Operational Comparison

Table 8: Four-corpus pooled results under mixed Medium-strength attacks ($n = 4,288$); these denominators differ from the public-only and poisoning-only tables.

Method	TCV↑	TAA↑	ASR↓	AR	P50 (ms)
Plain RAG	28.4 ± 1.2	34.2 ± 1.4	72.8 ± 1.6	0.0	234
Recency	46.2 ± 1.8	52.6 ± 1.6	51.4 ± 2.2	0.0	240
Per-Doc Latest	55.8 ± 1.4	60.4 ± 1.8	42.6 ± 1.8	0.0	246
Time Filter	67.4 ± 1.6	72.8 ± 1.4	34.8 ± 2.0	1.8	256
Agentic-RAG	74.2 ± 2.4	78.6 ± 2.2	24.6 ± 2.8	3.2	892
Multi-Source	81.6 ± 1.8	82.4 ± 2.0	19.4 ± 2.4	12.6	312
RevisionScope	93.2 ± 0.8	90.6 ± 1.0	8.6 ± 1.2	5.4	278

Table 8 extends the public result to policy and incident-response corpora. Explicit filtering supplies the largest heuristic gain; Agentic-RAG pays 892 ms P50 for 24.6% ASR; and Multi-Source reaches 19.4% ASR with 12.6% abstention. RevisionScope provides the strongest pooled safety–utility point at 278 ms.

A.2 Per-Dataset Results

Table 9 reports the public and operational corpora separately. The public rows reproduce the pooled ordering, while policy and incident-response rows test the same mechanism on different document types. Pooled estimates and intervals are computed directly over all question clusters in Table 8.

Table 9: Per-dataset results under mixed Medium-strength attacks.

	RevisionScope				Plain RAG			
Dataset	TCV	ASR	AR	P50	TCV	ASR	AR	P50
SecAdvisory	92.4 ± 1.2	9.2 ± 1.4	5.8 ± 0.8	276	27.8 ± 2.4	73.4 ± 2.8	0.0	232
PolicyDocs	94.6 ± 0.8	7.8 ± 1.0	7.8 ± 1.2	284	29.6 ± 2.0	71.8 ± 2.4	0.0	236
IncidentRunbooks	93.8 ± 0.6	8.2 ± 0.8	4.2 ± 0.6	272	28.2 ± 1.8	72.4 ± 2.2	0.0	234
CVEFixes-DB	92.2 ± 1.4	9.6 ± 1.6	5.6 ± 1.0	280	27.4 ± 2.6	74.2 ± 3.0	0.0	238

Table 9 shows consistent effects across all four corpora. RevisionScope stays between 92.2% and 94.6% TCV, while Plain RAG stays between 27.4% and 29.6%; latency varies by at most 12 ms within either method. Policy and incident-response rows preserve the ordering observed on the public corpora.

A.3 Baselines and Equal-Budget Comparison

Table 10 fixes every baseline interface and budget before comparison. The family spans recency, hard time filtering, extra-compute re-querying, source corroboration, factual citation verification, and learned source reliability. Table 12 then isolates the effect of targeted temporal controls from added inference cost.

All rows receive the same frozen root request and context, Contriever candidate pool, and 4k-token generation budget. Non-Agentic rows use Llama-2-7B-Chat; Agentic-RAG uses the specified GPT-4o-mini controller for up to three retrieval rounds. The Provenance adaptation applies its factual-support and citation-entailment decision after the shared retrieval pass. The RA-RAG adaptation combines source-reliability estimates with relevance and answer-support scores before selecting the shared $k = 10$ budget. Both use their frozen calibrated reject decisions for abstention while withholding $t_q^\star$, applicability intervals, and revision-validity labels.

Table 12 shows that generic extra compute does not reproduce the gain. Budget-matched Agentic-RAG reaches 31.4% ASR, and Multi-Source spends 12.6% abstention to reach 19.4% ASR. Revision-

Table 10: Baseline specifications.

Baseline	Configuration
Plain RAG	Contriever top-k=10, no temporal filtering, Llama-2-7B-Chat generation
Recency	Rerank retrieved passages by `last_updated` within top-k
Per-Doc Latest	Keep only the newest revision per document id
Time Filter	Hard filter with $t_d \leq \hat{t}_q$ and a 365-day look-back window
Agentic-RAG	GPT-4o-mini (2024-07-18), at most 3 rounds, 2,048 tokens per round, explicit freshness follow-up query
Multi-Source	Require agreement from at least two independent vendor sources; abstain otherwise
Provenance	Post-retrieval factual and citation-support verifier adapted from Sankararaman et al. (2024); one verifier pass, no revision-validity labels
RA-RAG	Reliability-weighted reranking adapted from Hwang et al. (2025); public source metadata and support scores, no revision-validity labels
RevisionScope	Root-scope resolution + epoch-bound cache + time-conditioned retrieval ($k = 10$, 4k context) + DeBERTa verifier ($\tau_{\text{hi}} = 0.65$, $\tau_{\text{lo}} = 0.35$) + conflict-aware generation

Table 11: Serving budget for the closest public comparators. Calls count retrieval passes; P95 is end-to-end latency on the 2,144-query public attacked slice.

Method	Retrieval calls	Verifier passes	Context	P95 (ms)
Provenance	1	1	4k	658
RA-RAG	1	1	4k	706
RevisionScope	1	1	4k	612

Table 12: Equal-budget comparison on the four-corpus pooled evaluation ($n = 4,288$, mixed Medium-strength attacks).

Method	TCV↑	ASR↓	AR	P50 (ms)
Agentic-RAG (unconstrained)	74.2	24.6	3.2	892
Agentic-RAG (equal-budget)	68.4	31.4	2.8	486
Multi-Source	81.6	19.4	12.6	312
RevisionScope	93.2	8.6	5.4	278

Scope reaches 8.6% ASR with 5.4% abstention and lower P50, locating the improvement in temporal interfaces rather than inference volume.

A.4 Scaling, Runtime, and Attack Strength

Table 13: Scaling grid for RevisionScope under mixed Medium-strength attacks.

Setting	TCV	ASR	P95 (ms)	QPS
k=5 (4k tokens)	91.4	10.2	486	3.98
k=10 (4k tokens)	93.2	8.6	612	3.62
k=15 (4k tokens)	94.2	8.0	714	3.28
k=20 (4k tokens)	94.8	7.8	816	3.05
2k tokens (k=10)	92.0	9.4	486	4.28
4k tokens (k=10)	93.2	8.6	612	3.62
6k tokens (k=10)	93.9	8.2	714	3.04
8k tokens (k=10)	94.4	8.0	816	2.41

Table 13 identifies $k = 10$ with a 4k-token context as the serving knee point. Raising k from 10 to 20 improves TCV by 1.6 points while adding 204 ms P95 latency; expanding context from 4k to 8k gains 1.2 TCV points while reducing throughput from 3.62 to 2.41 QPS. The default therefore captures most safety gain before runtime cost accelerates.

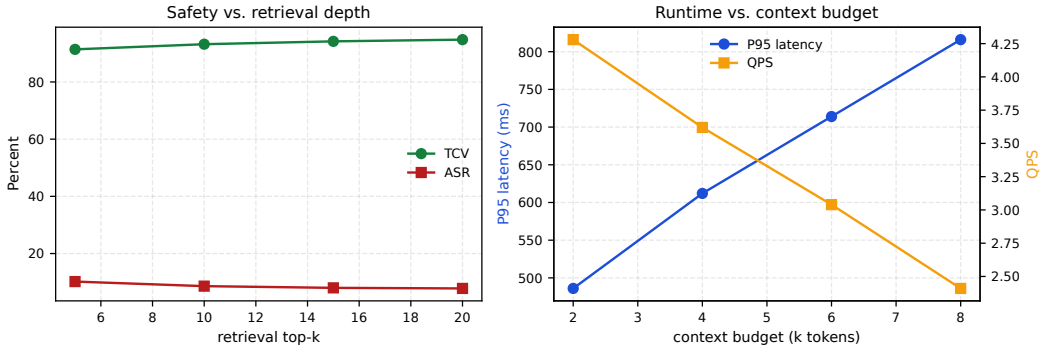

Figure 3: Scaling trade-offs for retrieval depth and context length.

Figure 3 makes the saturation pattern easier to read than the grid alone. Safety improves steadily as retrieval depth increases, but the slope is modest after the default operating point. Runtime cost, in contrast, rises almost linearly with context budget. The default captures most of the available safety gain before additional context becomes expensive and only marginally safer, locating targeted temporal controls on a more efficient frontier than brute-force scaling.

Figure 6 maps the two most distinct interventions to application behavior. The plain stack accepts shifted scope in panel (a) and a current but poisoned revision in panel (b); RevisionScope responds through task-scope canonicalization and conflict exposure. Together with Table 13 and Figures 3–5, the result separates targeted temporal controls from additional retrieval depth, context, or inference cost.

A.5 Generalization Across Models and Retrievers

Table 14 preserves a 64.4–65.6 point TCV advantage across four generators and a 64.8–66.6 point advantage across four retrievers. DPR improves both Plain RAG and RevisionScope, while the between-system separation remains, showing that the control effect is complementary to retriever quality. The same ordering holds across the tested open backbones, instruction-tuned variants, sparse retrieval, dense retrieval, and late interaction.

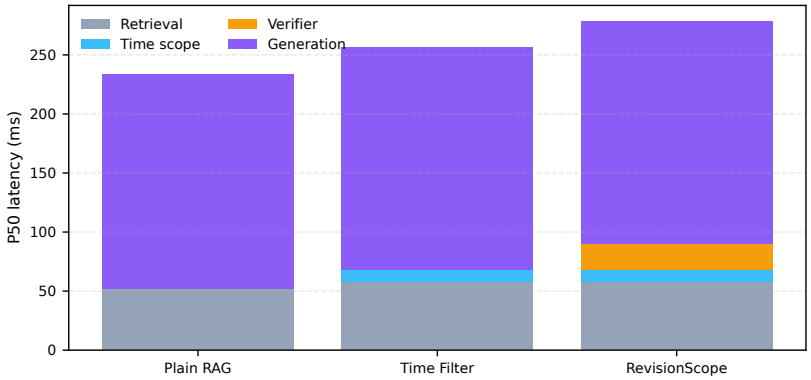

Figure 4: P50 latency decomposition.

Interpretation. Revision-aware retrieval and time-scope handling add 16 ms, applicability verification adds 22 ms, and generation remains 188 ms. The 44 ms total overhead matches the single-pass budget in Table 11.

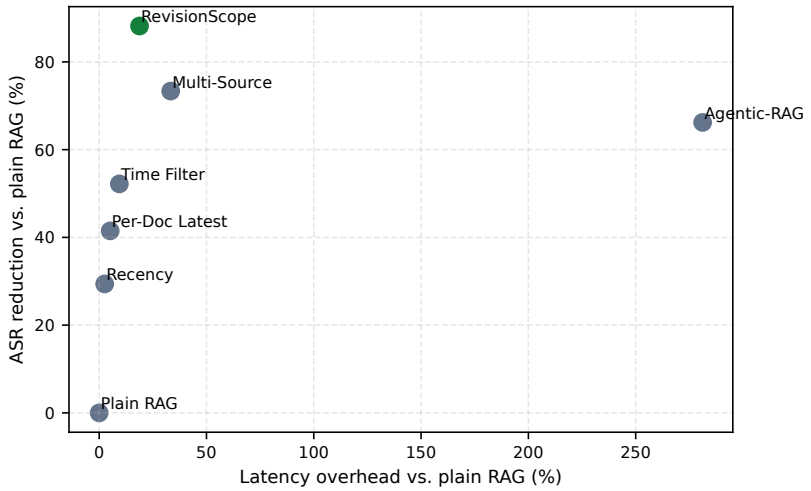

Figure 5: ASR reduction vs. latency overhead.

Interpretation. Agentic-RAG spends 281% P50 overhead for a 66% relative ASR reduction, Multi-Source trades more abstention for 73%, and RevisionScope reaches 88% at 19% P50 overhead.

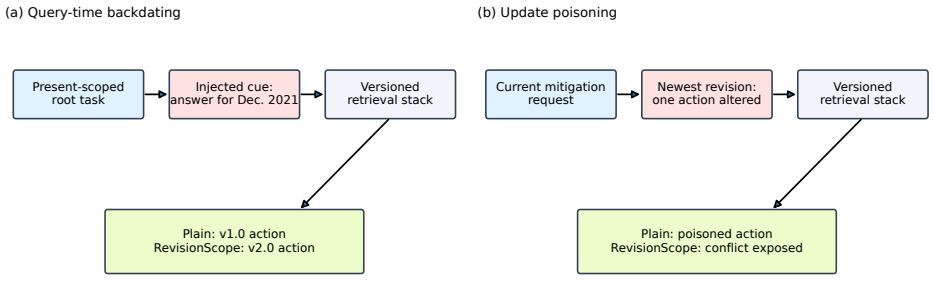

Figure 6: Schematic intervention paths: plain retrieval follows a backdated or poisoned revision, while RevisionScope routes by intended scope or exposes an action conflict.

Table 14: Cross-model and cross-retriever generalization under mixed Medium-strength attacks. The upper block varies the generator with Contriever fixed; the lower block varies the retriever with Llama-2-7B-Chat fixed.

	RevisionScope		Plain RAG	
LLM	TCV	ASR	TCV	ASR
Llama-2-7B-Chat	93.2	8.6	28.4	72.8
Llama-3-8B-Instruct	94.6	7.8	30.2	69.4
Mistral-7B-Instruct-v0.3	92.4	9.4	26.8	74.2
Phi-3-mini-4k-instruct	90.8	11.2	25.4	76.8
Retriever				
BM25	91.2	10.4	25.8	75.6
Contriever	93.2	8.6	28.4	72.8
DPR	94.4	7.8	29.6	71.2
ColBERT-v2	93.8	8.2	27.2	73.4

B Supplementary Analysis and Failure Cases

B.1 Component Ablation

Table 15: Component ablation for RevisionScope.

Configuration	TCV↑	TAA↑	ASR↓	ΔASR
Full RevisionScope	93.2	90.6	8.6	—
−Conflict-Aware Gen.	85.6	82.4	17.4	+8.8
−Temporal Verifier	72.4	68.8	31.8	+23.2
−Time-Cond. Retrieval	59.8	56.2	46.2	+37.6
−All (= Plain RAG)	28.4	34.2	72.8	+64.2

Table 15 explains why the core stack is ordered the way it is. Removing time-conditioned retrieval causes the largest regression because it prevents the rest of the system from starting from a temporally valid candidate set. Removing the verifier also hurts substantially, which shows that timestamp rules alone are not enough once evidence is ambiguous or inconsistently marked. Conflict-aware generation still matters because some disagreement is only visible after retrieval and verification have done their work. The ablation therefore identifies RevisionScope as a threat-aligned composition whose interfaces contribute in the order predicted by the attack model.

B.2 Verifier Calibration Behavior

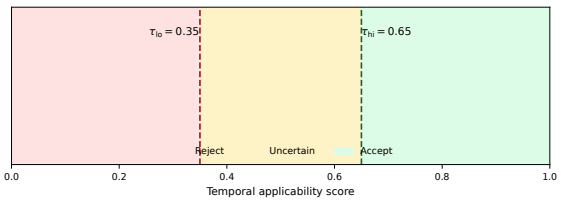

Figure 7: Verifier triage on the native 0–1 score scale. Shaded regions encode reject (< 0.35), uncertain (0.35–0.65), and accept (≥ 0.65) states.

Figure 7 makes threshold units explicit and shows the state passed downstream. Scores in the middle band preserve uncertainty for conflict-aware generation instead of forcing a binary retrieval decision; the cross-family calibration row in Table 7 measures the cost of adapting these boundaries.

B.3 Poisoning Mitigation Extension

Action-level validation extends the core system with the correction-or-abstention pass defined in Section 4. All metrics score the final returned action and citations. The extension reduces poisoning ASR most strongly on partial-truth cases while exposing the corresponding abstention and latency cost.

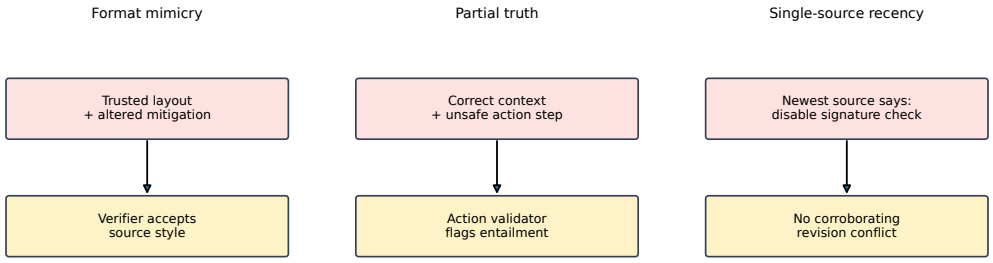

Figure 8: Schematic mechanisms targeted by the poisoning evaluation and action-validation extension.

Figure 8 separates three poisoning mechanisms. Format mimicry preserves source style, partial truth alters one operational step inside otherwise correct context, and single-source recency removes the cross-revision disagreement signal. These cases motivate source-reliability comparators, action validation, and the adaptive variants in Table 21.

Table 16: Action-level validation on the four-corpus pooled poisoning subset ($n = 1,072$). TCV and ASR score final outputs after correction or abstention; ASR uses the poisoning-only denominator.

Configuration	TCV	ASR	AR	ΔASR
RevisionScope (baseline)	93.2	15.8	5.4	—
+ Action-level validation	94.6	10.2	7.2	−5.6
By poisoning failure mode after action-level validation				
Format mimicry	91.8	14.6	8.4	—
Partial truth	96.2	6.8	6.8	—
Single-source recency	95.8	8.4	6.2	—

Table 16 uses the 1,072-query four-corpus poisoning subset. Action validation lowers ASR from 15.8% to 10.2% and raises AR from 5.4% to 7.2%. The largest gain appears on partial truth, where imperative-step entailment directly targets the corrupted action; format mimicry has the highest residual ASR.

C Public Evidence Extensions

C.1 Public Answer Audit and Raw Counts

Table 17 evaluates the answer-level judge on a stratified public subset drawn after model outputs are frozen. It covers Plain RAG, Time Filter, Multi-Source, and RevisionScope across all attacks and both abstaining and non-abstaining traces. With timestamps and revision identifiers hidden, the same 240 traces yield 93.3 [3.2]% judge–human agreement for the CS label used in Table 4. Disagreements concentrate in abstention-boundary cases (11), mixed-citation traces (7), and residual label ambiguity (6); none changes the TCV ordering.

Table 17 reconciles every agreement percentage with its raw disagreement count. The answer-level judge preserves the method ordering, and timestamp-hidden CS agreement remains high, validating the content-only side of the support–scope comparison. Disagreements concentrate in abstention-

Table 17: Human audit of TAA/ASR on public attacked traces. Judge–human agreement is $1 - \text{Disag.}/n$; unsafe precision and recall use adjudicated labels.

Slice	n	Human-Human	Judge-Human	Unsafe P	Unsafe R	Disag.
Overall	240	93.7 [3.0]	92.5	89.8 [4.2]	87.5 [4.8]	18
Plain RAG	60	—	91.7	92.6 [4.1]	95.1 [3.6]	5
Time Filter	60	—	90.0	86.9 [6.3]	84.2 [7.1]	6
Multi-Source	60	—	93.3	88.1 [6.0]	85.7 [6.8]	4
RevisionScope	60	—	95.0	89.4 [5.8]	83.6 [7.0]	3

boundary and mixed-citation cases rather than obvious unsafe traces. The public temporal-validity claim remains anchored in judge-free TCV.

Table 18: Public per-attack successes over 1,608 query-runs per column (536 frozen questions $\times$ three seeds). Cells show raw successes followed by ASR and bracketed 95% bootstrap half-width.

Method	Backdate	Spoof	Poison	Replay
Plain RAG	1,101; 68.5 [2.3]	1,192; 74.1 [2.1]	1,319; 82.0 [1.9]	1,135; 70.6 [2.2]
Time Filter	461; 28.7 [1.8]	569; 35.4 [2.0]	669; 41.6 [2.1]	397; 24.7 [1.7]
RA-RAG	195; 12.1 [1.4]	257; 16.0 [1.6]	304; 18.9 [1.7]	196; 12.2 [1.4]
RevisionScope	125; 7.8 [1.1]	170; 10.6 [1.3]	204; 12.7 [1.4]	105; 6.5 [1.0]

Equal public denominators make the mechanism comparison direct: the average of each row matches its overall ASR in Table 5. RA-RAG is the strongest generic baseline on every attack, while RevisionScope reduces raw successes further across all four channels and abstains less overall. Poison remains the hardest column for both methods, motivating the adaptive analysis below.

C.2 Held-Out Validation and Public Poisoning

Table 19 evaluates a separately authored public set whose prompts, thresholds, and baselines were frozen before inference. It contains 256 SecAdvisory and 256 CVEFixes-DB questions. The milder attack rates quantify distribution shift away from the construction set while preserving a large temporal-integrity gap.

Table 19: Held-out temporal-integrity validation on a same-domain, separately authored public query set; cells report the mean with bracketed 95% bootstrap half-width over three seeds.

Method	TCV↑	ASR↓	AR
Plain RAG	41.8 [2.1]	46.9 [2.8]	0.0
Time Filter	74.6 [1.8]	25.4 [2.1]	1.2
Multi-Source	86.1 [1.4]	14.8 [1.9]	10.7
RevisionScope	94.8 [0.9]	6.1 [1.0]	4.7

Across 1,536 held-out query-runs, Plain RAG produces 642 valid-citation and 720 unsafe traces; RevisionScope produces 1,456 valid-citation and 94 unsafe traces. Time Filter records 1,146/390, and Multi-Source records 1,322/227. These counts align the aggregate intervals with inspectable outcomes and confirm that the ranking persists under independently written prompts.

Table 20 isolates 384 public poisoning-compatible questions: 128 format-mimicry, 128 partial-truth, and 128 single-source-recency cases. Provenance, RA-RAG, and Multi-Source test factual verification, source reliability, and corroboration against the core temporal controls and action-validation extension.

Over 1,152 query-runs, raw valid/unsafe counts are 264 / 917 for Plain RAG, 574 / 661 for Time Filter, 847 / 362 for Provenance, 889 / 309 for Multi-Source, 952 / 264 for RA-RAG, 1,033 / 212 for RevisionScope, and 1,049 / 134 with action validation. RA-RAG is the strongest generic comparator,

Table 20: Public poisoning-compatible subset ($n = 384$); metrics score final outputs after any correction-or-abstention pass. Brackets are 95% bootstrap half-widths over three seeds; ASR uses a poisoning-only denominator.

Method	TCV↑	ASR↓	AR	P50 (ms)
Plain RAG	22.9 [2.4]	79.6 [3.0]	0.0	236
Time Filter	49.8 [2.1]	57.4 [2.6]	1.4	255
Provenance	73.5 [1.9]	31.4 [2.3]	8.8	308
Multi-Source	77.2 [1.8]	26.8 [2.2]	14.1	319
RA-RAG	82.6 [1.6]	22.9 [2.0]	10.4	333
RevisionScope	89.7 [1.3]	18.4 [1.8]	5.8	281
RevisionScope + action-level validation	91.1 [1.2]	11.6 [1.5]	7.6	337

yet the action-aligned stack cuts its ASR nearly in half while abstaining less. Action validation supplies a further 6.8-point ASR reduction at a 56 ms P50 cost, matching its partial-truth design target.

C.3 Adaptive Verifier and Conflict Attacks

Adaptive templates are developed on examples excluded from the final set and target disclosed thresholds, partial-truth actions, source-recency weights, or conflict suppression. The final evaluation freezes 96 questions per attack and reports successes over 288 query-runs. Every defense receives the same candidate pool and attack text; action-validation rows add the single correction-or-abstention pass and are scored on its final output.

Table 21: Adaptive temporal attacks against verifier and conflict logic. Brackets are 95% bootstrap half-widths over three seeds; raw successes use 288 query-runs per attack–method pair.

Adaptive attack	Method	Successes	TCV	ASR	AR	P50
Threshold language	Plain RAG	249	19.8 [2.8]	86.5 [2.6]	0.0	238
	Time Filter	190	44.2 [2.5]	66.0 [2.9]	1.7	258
	Provenance	127	65.4 [2.2]	44.1 [3.1]	9.6	312
	Multi-Source	108	71.2 [2.0]	37.5 [2.9]	15.3	324
	RA-RAG	95	77.6 [1.9]	33.0 [2.7]	11.8	337
	RevisionScope	82	83.1 [1.6]	28.5 [2.4]	7.2	286
	RevisionScope + action	60	86.4 [1.5]	20.8 [2.1]	9.8	343
Partial-truth action	Plain RAG	254	18.6 [2.9]	88.2 [2.4]	0.0	239
	Time Filter	196	42.5 [2.6]	68.1 [2.8]	1.4	259
	Provenance	133	63.8 [2.3]	46.2 [3.0]	8.6	313
	Multi-Source	115	69.6 [2.1]	39.9 [2.9]	13.8	325
	RA-RAG	103	75.1 [2.0]	35.8 [2.6]	11.2	338
	RevisionScope	97	80.8 [1.8]	33.7 [2.7]	6.8	287
	RevisionScope + action	47	88.7 [1.4]	16.3 [1.9]	10.6	344
Source recency	Plain RAG	242	22.7 [2.7]	84.0 [2.7]	0.0	237
	Time Filter	178	48.9 [2.4]	61.8 [3.0]	2.1	257
	Provenance	113	69.7 [2.1]	39.2 [2.9]	10.8	311
	Multi-Source	94	75.4 [1.9]	32.6 [2.7]	16.1	323
	RA-RAG	81	81.4 [1.7]	28.1 [2.5]	12.5	336
	RevisionScope	69	86.2 [1.5]	24.0 [2.2]	7.6	285
	RevisionScope + action	51	89.4 [1.3]	17.7 [1.9]	9.2	342
Conflict suppression	Plain RAG	247	20.9 [2.8]	85.8 [2.6]	0.0	238
	Time Filter	187	46.1 [2.5]	64.9 [2.9]	1.8	258
	Provenance	121	67.1 [2.2]	42.0 [3.0]	9.9	312
	Multi-Source	101	73.2 [2.0]	35.1 [2.8]	14.9	324
	RA-RAG	92	78.5 [1.9]	31.9 [2.6]	12.0	337
	RevisionScope	88	82.4 [1.7]	30.6 [2.5]	8.4	286
	RevisionScope + action	66	85.6 [1.6]	22.9 [2.2]	11.7	343

Interpretation. All defenses degrade under adaptive knowledge, but their ordering remains stable. RA-RAG is the strongest generic row; RevisionScope lowers its ASR with lower latency and abstention, and action validation is especially effective against partial-truth edits. The remaining ASR quantifies the operating boundary under disclosed verifier and conflict logic.

C.4 Security Evidence Matrix

Table 22 binds each threat capability to its protected object, trust boundary, measured property, closest comparator, adaptive or transfer check, benign cost, residual condition, and paper-internal evidence.

Table 22: Evidence matrix for temporal-integrity claims.

Attack / capability	Protected object; trusted boundary	Tested property; enforcement; comparator
Backdate / query adversary shifts resolved scope	Revision valid for $t_q^\star$; trust root/context boundary and revision history	Valid citation and scoped action; enforce at resolver and grouped retrieval; compare RA-RAG and Multi-Source
Spoof / metadata adversary alters time fields	Estimated applicability $\widetilde{I}(d)$; trust stable identity, immutable revision order, and revision text	Exclude stale evidence under misleading metadata; enforce at applicability triage; compare RA-RAG and Multi-Source
Replay / serving adversary returns stale cache	Snapshot valid for present scope; trust request epoch and revision chain	Reject epoch mismatch and revalidate surviving payload; enforce at cache admission and frontier construction; compare RA-RAG and Multi-Source
Poison / corpus adversary inserts malicious update	Actionable guidance in newest revision; trust task scope, identity map, and conflict policy	Preserve action consistency; enforce at conflict and action checks; compare RA-RAG and Multi-Source

Attack	Adaptive / transfer check; benign cost	Residual condition; evidence trace
Backdate	Held-out public, per-attack interval, and cross-family transfer; 271 ms P50, 1.6% AR	Ambiguous root scope; Tables 4, 18, 7; Figure 6a
Spoof	Held-out public, threshold-language attack, and cross-family transfer; 271 ms P50, 1.6% AR	Conflicting effective dates; Tables 5, 21, 7
Replay	Held-out public, source-recency attack, and cross-family transfer; 271 ms P50, 1.6% AR	Same-key stale payload or uncertain applicability; Tables 18, 21
Poison	Public poisoning and four adaptive variants; 337 ms P50, 7.6% AR with action validation	Partial truth and conflict suppression; Tables 16, 20, 21; Figures 6b, 8

Interpretation. Every attack changes one untrusted channel, names the boundary that remains fixed, and links the surviving failure to score-bearing evidence. RA-RAG and Multi-Source define the strongest generic comparator frontier; the remaining columns locate transfer, benign cost, and adaptive residual risk.