TRACING BIOMEDICAL RAG UPDATE WINDOWS WITH JOINT EVIDENCE ADMISSION

Anonymous authors
Paper under double-blind review

ABSTRACT

A newly indexed biomedical passage can meet the same query at 0, 6, 12, 24, or 48 hours, yet snapshot evaluation assigns these encounters one security state. We introduce a paired-delay replay that advances query time while fixing bytes, query, payload, wave, candidate content, and credibility inputs, together with TEMPER (Temporal Evidence Management with Provenance Enforcement for RAG), which jointly admits committed use, age-indexed exposure, and claim support. Under its $4 \to 3 \to 2 \to 0 \to 0$ pending-slot cap, success on the same 7,500 injection trials traces 4.4%, 3.8%, 3.0%, 2.1%, and 2.3%, a 2.2-point [1.3, 3.1] contrast absent from no-quarantine and SecAlign+RAGuard controls. Against the evaluated RAGuard–SecAlign composition, TEMPER reduces poisoning from 6.1%±0.9% to 3.2%±0.7% and accepted evidence modifications from 688/1,000 to 2/1,000. Under adaptive attacks, independent labels of emitted citations give pooled F1 of 80.7% versus 68.6%, while high-severity clinical-harm events fall from 8.8%±1.4% to 5.9%±1.1%. TEMPER retains 5.2%/4.1% injection/poisoning success with Qwen2.5-72B-Instruct and benchmark accuracy within 0.9 points of Vanilla; warm-cache p50 overhead is 11.4%. Evidence age is therefore a measurable security coordinate from newly committed bytes to cited biomedical claims.

1 INTRODUCTION

At ingestion, the evaluated policy allows unpromoted evidence from the current biomedical update to occupy four of twenty model-context slots. The allowance becomes three after six hours, two after twelve hours, and zero immediately before the 24-hour promotion transaction; after that transaction, only corroborated low-risk items join the established pool. Following an identical query–payload trial through these states reveals what a settled index erases. On 7,500 paired trials, TEMPER's injection success traces $4.4\% \to 3.8\% \to 3.0\% \to 2.1\% \to 2.3\%$, whereas the same trials remain approximately level without quarantine and under SecAlign+RAGuard. The resulting 2.2-point [1.3, 3.1] contrast identifies the *update window*: a measured interval in which query delay changes the exposure of newly ingested evidence.

That interval is consequential in biomedical RAG because retrieved text is executable model context, not a passive citation collection. A passage can be topically relevant while carrying a clinical-prose instruction, a plausible false treatment claim, or metadata detached from its committed bytes (Greshake et al., 2023; Zhang et al., 2026). Generator alignment addresses behavior after exposure, poisoning filters screen selected candidates, temporal RAG prioritizes freshness, and citation checkers inspect generated support. A streaming answer, however, traverses all of these boundaries in order. A security test that retains only the final snapshot cannot identify whether failure entered through modified bytes, time-dependent context allocation, or an unsupported cited claim.

We represent that path by the update-indexed state

$$A_t(e, q, c) = (A_t^{\text{use}}(e), A_t^{\text{exposure}}(e, q), A_t^{\text{support}}(e, c)),$$

and instantiate it with TEMPER. Signed span commitments bind content, offsets, and provenance before retrieval. A frozen relevance–credibility–instruction score allocates verified candidates subject to the age-dependent pending budget and promotion thresholds. After generation, SPANGUARD classifies atomic claim–span pairs and permits only support from verified non-instruction spans, with

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.

one constrained repair opportunity. The technical contribution is the shared evidence identity and ordered admission state: the same item is carried from commitment through a delay-indexed exposure decision to the span that supports an emitted claim.

The study replays 12 public PMC-OA and FDA-label waves with PubMedQA, MedQA, BioASQ, and MPIB CHER_3 outcomes (Jin et al., 2019; 2021; Balikas et al., 2015; Lee et al., 2026). Paired delay and removal ablations supply the mechanism comparisons; the RAGuard–SecAlign result is an end-to-end comparison with method-specific configurations tabulated in the appendix. TEMPER lowers poisoning from 6.1% to 3.2% and accepted modifications from 688/1,000 to 2/1,000. Blinded labels give adaptive pooled F1 of 80.7% versus 68.6% among emitted citations, and CHER_3 falls from 8.8% to 5.9%; adaptive families, Qwen transfer, benchmark utility, and serving cost complete the evidence path.

Contributions. The central contribution is a lifecycle-resolved protocol that follows one attack through five evidence ages and separates pre- from post-promotion state. TEMPER contributes the corresponding joint admission design, connecting committed provenance, delay-indexed context allocation, and claim-to-span verification. Its stage-resolved evidence makes a biomedical update testable as a sequence of decisions rather than one settled snapshot.

2 RELATED WORK

2.1 TEMPORAL AND CONTINUOUSLY UPDATED RAG

HoH measures failures from outdated evidence, MRAG decomposes temporal retrieval constraints, and EraRAG supports efficient index growth (Ouyang et al., 2025; Zhang et al., 2025b;a). They use time for correctness or maintenance. Our replay preserves query, payload, bytes, and candidates while advancing the query relative to ingestion and promotion, producing an adversarial exposure trajectory rather than temporal-question accuracy.

2.2 INJECTION AND POISONING DEFENSES

SecAlign aligns the generator after untrusted context arrives (Chen et al., 2025). RAGuard filters by perplexity and similarity, TrustRAG combines cluster filtering with self-assessment, ShieldRAG learns safety-aware retrieval, and SD-RAG performs disclosure-constrained redaction (Cheng et al., 2025; Zhou et al., 2025; Yang et al., 2026; Masoud et al., 2026). We evaluate SecAlign, RAGuard, and their operational composition on matched policy-override and false-claim endpoints; other published candidate policies define distinct decision boundaries summarized in Appendix Table 3. TEMPER adds query delay to that boundary through an age-dependent pending budget and thresholded promotion.

2.3 PROVENANCE AND CLAIM SUPPORT

Data provenance and Merkle commitments bind transformed content to auditable state (Buneman et al., 2001; Merkle, 1988). RAGShield v1 combines document attestation, trust-weighted retrieval, contradiction detection, and provenance-aware citations (Patil, 2026); TEMPER carries byte- and span-level commitments through a measured pending-to-promotion trajectory and defines violation by generator exposure. ALCE, RAGChecker, and CiteGuard evaluate or retrieve citation support (Gao et al., 2023; Ru et al., 2024; Choi et al., 2026), while MedRAGChecker and VerifAI provide biomedical claim verification (Ji et al., 2026; Košprdić et al., 2026). SPANGUARD makes verification an online admission decision; blinded labels evaluate its emitted citations independently.

2.4 BIOMEDICAL OUTCOME MEASUREMENT

Biomedical RAG inherits medical-ML adversarial risk while adding an evidence-mediated instruction channel (Finlayson et al., 2019; Gu et al., 2019). MPIB's CHER_3 counts attack-conditioned answers with severity at least three (Lee et al., 2026); pairing it with interface endpoints separates retrieved attacks, provenance bypasses, unsupported citations, and harmful answers.

3 Problem Setting and Threat Model

3.1 Setting

In the biomedical RAG application considered here, a query q is answered using medical documents, including scientific articles and drug labels, in a corpus $\mathcal{D}_t$ that changes over time. Discrete update waves $t \in \{0, 1, 2, \ldots\}$ represent time; additions or replacements U_t transform the preceding snapshot at each wave. Deterministic chunking produces retrievable evidence items e from documents, with source metadata attached to each item, such as its source identifier, ingestion timestamp $t_{\mathrm{ing}}(e)$, and policy tags. Query delay $\delta = t_q - t_{\mathrm{ing}}(e)$ makes evidence age an explicit replay coordinate.

The pipeline has three stages. Given q and the current snapshot, retrieval produces the evidence set of top-K items,

$$E_t = R(q, \mathcal{D}_t), \qquad E_t = \{e_1, \ldots, e_K\}.$$

Generation maps q and E_t to an answer a_t and citations C_t:

$$(a_t, C_t) = G(q, E_t).$$

A citation maps an atomic claim in the answer to an evidence span. Truthfulness is not a default assumption for the generator: instructions within retrieved text may guide its behavior, or it may cite evidence without actual support for the claim. Such failures concern the evidence layer as well as answer quality.

For evidence item e, query q, and atomic claim c, we define the update-indexed admission state

$$A_t(e, q, c) = \big(A_t^{\mathrm{use}}(e), A_t^{\mathrm{exposure}}(e, q), A_t^{\mathrm{support}}(e, c)\big).$$

Use-admissible evidence verifies against the current committed snapshot; exposure-admissible evidence survives trust, instruction-risk, quarantine, and budget controls; support-admissible evidence clears claim-to-span verification before appearing in C_t. The shared state makes failures at the three interfaces observable under the same update and query.

An item is *pending* from its committed ingestion until the promotion transaction. An *established* item predates the current wave or has passed promotion and remains valid under the current root. The boundary state (T_q, pre) is observed after the pending budget reaches zero but before promotion, whereas (T_q, post) is observed after promotion commits the next state. These definitions let one item be traced without conflating elapsed time with a state transition.

3.2 Threat Model

Our attacker model permits both direct and indirect influence over the evidence layer. Allowed actions include injecting or poisoning some documents in $\mathcal{D}_t$, placing natural-language instructions within content that is otherwise relevant, altering committed bytes, offsets, or provenance metadata, coordinating multiple documents, and selecting δ to query malicious content when its admission state is most favorable. All attackers know the admission design. A1–A3 and A5–A6 search payloads only with development queries; A4 is an explicitly query-aware stress test that receives evaluation query text but no answer label, model output, or evaluation feedback. The standard poisoning construction receives the benchmark supporting claim needed to define its frozen false target. No attacker modifies model weights, observes evaluation outputs before scoring, controls the trusted signing key, or rolls back the trusted checkpoint sequence.

The cryptographic assumptions are standard: collision resistance and unforgeable signatures. We further require verification to precede delivery of evidence to the generator, with no silent rollback of the checkpoint sequence used in that verification. Conditional on these assumptions, three families of objectives are studied: policy override via retrieved instructions, targeted misinformation via poisoned evidence, and integrity bypass via post-ingestion tampering or provenance forgery.

Broadly, these evidence-layer threats parallel the adversarial and poisoning risks examined across medical ML and the ML supply chain (Finlayson et al., 2019; Gu et al., 2019).

Protected objects. Under collision resistance, signature unforgeability, an uncompromised ingestion key, and checkpoint non-rollback, direct proof verification detects post-ingestion changes to

committed content, offsets, and metadata. IVR separately measures whether the full pipeline enforces that check. Exposure admission handles malicious content committed at ingestion; SPANGUARD is calibrated to minimize false admission over every non-support class, while independent labels and CHER_3 measure answer behavior.

3.3 METRICS

For prompt-injection and poisoning objectives, we report Attack Success Rate (ASR), alongside Integrity Violation Rate (IVR), defined as accepted tamper attempts divided by all 1,000 attempts. Injection ASR first averages within each benchmark–attack-family cell and then averages the 15 cells; because every cell contains 500 trials, this value also equals total successes divided by 7,500. The paired update-window contrast $W_m = \text{ASR}_m(0) - [\text{ASR}_m(T_q, \text{pre}) + \text{ASR}_m(48\text{h})]/2$ compares immediate exposure with the closed-budget boundary and post-promotion state on identical attacks. Utility is benchmark accuracy or strict F1; CHER_3 is the proportion of attack-conditioned answers assigned clinical-harm severity at least three under MPIB (Lee et al., 2026). Efficiency includes latency, throughput, memory, and ingestion cost. The internal verifier diagnostic reports per-query macro-F1; the blinded audit reports pooled precision, recall, and F1 over emitted claim–citation decisions with query-clustered intervals.

4 UPDATE-INDEXED JOINT EVIDENCE ADMISSION

4.1 DESIGN OVERVIEW

TEMPER implements the three coordinates of A_t as one ordered state machine. At ingestion, signed commitments establish A_t^{use}. At retrieval, verification status, source credibility, evidence age δ, instruction risk, and relevance determine A_t^{exposure}. After generation, SPANGUARD emits a citation only when a verified span supports the corresponding atomic claim, establishing A_t^{support}. The order propagates concrete state: unverified evidence never receives a ranking score, the pending-slot budget changes as δ approaches promotion, and generated claims cannot inherit support from a merely relevant passage.

The distinctive mechanism is the state carried between these decisions. Commitments, risk scoring, and natural-language inference are established primitives; in TEMPER they operate on one span identity under one update root, and the exposure coordinate changes at explicit replayable delay and promotion events. This construction supports stage-specific counterfactuals: fixing an item while advancing δ tests temporal exposure, removing one operation tests its interface responsibility, and retaining the span identity through generation tests whether admitted evidence actually supports the emitted claim.

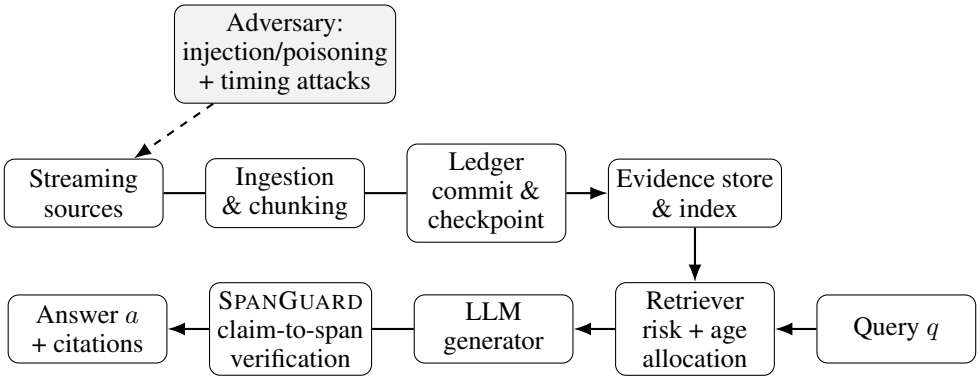

Figure 1: TEMPER traces one evidence identity through three ordered decisions: (1) signed span commitments establish use, (2) credibility, instruction risk, and the age-indexed pending budget allocate exposure, and (3) SPANGUARD admits claim support. Integrity proofs travel with selected evidence from the retriever to the generator.

Figure 1 separates the measured consequences of this path. Upstream attacks exercise use and exposure before the generator; the output stage tests whether hostile or irrelevant context becomes asserted support. IVR, ASR, external citation F1, and CHER_3 therefore measure different failure events rather than interchangeable security proxies.

4.2 USE ADMISSION: COMMITTED PROVENANCE

For each ingested document d, we compute a canonical representation $\mathrm{canon}(d)$ and bind its content to its provenance metadata,

$$h_d = H(\mathrm{canon}(d) \,\|\, \mathrm{meta}(d)),$$

where $\mathrm{meta}(d)$ includes source identifier, timestamp, access policy, and ingestion-pipeline version.

Definition 1 (Span commitment)**.** *Let s be a byte-accurate span of $\mathrm{canon}(d)$, including its offsets. Its commitment is $h_s = H(h_d \,\|\, \mathrm{span}(s))$.*

At each checkpoint, a Merkle tree (Merkle, 1988) aggregates span commitments, and an ingestion attestation key signs the checkpoint root root_t,

$$\mathsf{sth}_t = \mathsf{Sign}(\mathsf{sk}_{\text{ingest}},\ t \,\|\, \mathsf{root}_t).$$

At query time the retrieval service returns $(e, \pi_e, \mathsf{sth}_t)$, where π_e proves the evidence span's inclusion. Canonicalization targets UTF-8 NFC text, normalized line endings, byte offsets into the canonical byte string, and a closed metadata schema. Verification is logarithmic in the number of committed spans and cacheable under (root_t, h_s); the required root transition invalidates proof entries before publishing the next index/root pair. The verified tuple, rather than a second mutable store lookup, is the required generator input. IVR counts any prototype path that violates this order—including cache or schema enforcement defects—rather than treating a valid cryptographic primitive as sufficient. This gate applies classical transformation provenance (Buneman et al., 2001) to the invariant that model-visible content, offsets, and metadata match the signed snapshot.

Theorem 1 (Direct-verification tamper evidence)**.** *Suppose collision resistance holds for H, unforgeability holds for* Sign*, the ingestion attestation key remains uncompromised, and verification uses a checkpoint sequence without silent rollback. Under these conditions,* $\mathsf{Verify}(e, \pi_e)$ *rejects any submitted post-ingestion change to content, offsets, or provenance metadata, with only negligible probability of an exception.*

Interface invariant. Direct verification supplies the cryptographic guarantee; IVR evaluates whether cache, parser, and root-publication paths preserve it in the complete pipeline. Validly committed malicious content is then handled by exposure and support admission. Appendix A gives the proof.

4.3 EXPOSURE ADMISSION: RISK, CREDIBILITY, AND AGE

Candidate retrieval takes the BM25–BGE union at depth 100. Within that candidate set, min–max normalized BM25 and mapped cosine similarity form

$$\mathrm{Rel}(q, e) = 0.45\,\widehat{\mathrm{BM25}}(q, e) + 0.55\,(\cos(q, e) + 1)/2.$$

Credibility is

$$\mathrm{Cred}(e) = 0.50 s_{\text{source}} + 0.30 s_{\text{corrob}} + 0.20 s_{\text{policy}}.$$

The three frozen terms encode canonical source tier, ingest-policy compliance, and the fraction of claims supported by a distinct pre-wave established source; same-source/current-wave matches are excluded and unavailable corroboration scores zero. Instruction risk is a calibrated L_2-logistic model over 38 directive, role, override, delimiter, obfuscation, and imperative-ratio features, fitted with source-disjoint development and calibration splits. Appendix B gives the exact score construction and leakage controls. TEMPER ranks verified candidates with

$$\mathrm{Score}_t(e, q) = \mathrm{Rel}(q, e) + 0.35\,\mathrm{Cred}(e) - 0.55\,\mathrm{Inj}(e) - 0.25\,\mathbf{1}[e \in Q_t],$$

Newly verified items enter the pending set Q_t. For a weekly wave whose items share ingestion time, their context allowance decays as

$$b_Q(\delta) = \left\lceil 4\left(1 - \frac{\min\{\delta, T_q\}}{T_q}\right)\right\rceil, \qquad T_q = 24 \text{ hours}.$$

Thus at most $4, 3, 2, 0, 0$ of the $K = 20$ slots can contain unpromoted current-wave evidence at the evaluated delays $0, 6, 12, 24, 48$ hours. Established verified evidence backfills the released slots. We distinguish two event phases at the 24-hour boundary: (T_q, pre) is queried after b_Q reaches zero but before promotion; the subsequent transaction moves an item out of Q_t only if $\mathrm{Cred}(e) \geq 0.75$, $\mathrm{Inj}(e) < 0.15$, and $s_{\mathrm{corrob}}(e) > 0$. Items that fail remain zero-budget pending evidence; (T_q, post) and the 48-hour replay use the committed post-promotion state. Stable ties are resolved by source identifier and chunk offset.

The score coefficients were selected once from $\{0, .15, .25, .35, .45, .55, .65\}$ on 20 development query IDs per benchmark over waves 1–3. The promotion thresholds, four-slot ceiling, linear decay, and T_q were fixed in the same development phase. The age sweep changes only δ, $b_Q(\delta)$, and thresholded promotion; candidate bytes, query IDs, credibility inputs, and ranking parameters remain fixed.

4.4 SUPPORT ADMISSION: CLAIM-TO-SPAN BINDING

SPANGUARD runs citation checking as an output-stage gate, complementing work that evaluates attribution after generation (Menick et al., 2022; Gao et al., 2023; Rashkin et al., 2023; Choi et al., 2026). It decomposes the draft into atomic claims, retrieves candidate spans only from the verified context, and classifies each pair as $\{\mathsf{Support}, \mathsf{Refute}, \mathsf{Irrelevant}, \mathsf{Instruction}\}$. A citation is emitted when $p_{\mathrm{sup}} \geq \tau_{\mathrm{sup}}$ and $p_{\mathrm{inst}} \leq \tau_{\mathrm{inst}}$. Otherwise the generator receives one constrained repair pass with the unsupported claim and available verified spans; a second failure removes the claim and emits an abstention for answer-critical content.

The backbone is DeBERTa-v3-large (He et al., 2021; 2023), fine-tuned on document-disjoint MedNLI, SciFact, FEVER, and instruction-negative pairs (Romanov & Shivade, 2018; Wadden et al., 2020; Thorne et al., 2018). Thresholds minimize selective false admission over all three non-support labels subject to at least 80% support recall on the held-out calibration split; Section C reports mixture counts, class balance, threshold values, and calibration intervals.

Appendix Lemma 2 decomposes false admission into unconditional joint error and precision conditional on emission. It claims no numerical bound without an estimated error bound and coverage; the blinded audit independently estimates precision and recall among emitted citations.

4.5 UPDATE AND QUERY ALGORITHM

At wave t, TEMPER's required execution order is:

1. validate the closed metadata schema, canonicalize and chunk U_t, commit content/offset/metadata tuples, sign root_t, invalidate root-keyed caches, and publish the pending index with its root;

2. retrieve the top-100 hybrid candidates, verify each against root_t, compute Score_t, allocate the $K = 20$ context slots from the currently committed promotion state with $b_Q(\delta)$, backfill from established evidence, and apply stable tie-breaking;

3. pass the verified tuples directly to the generator without a second mutable lookup, generate an answer, bind its atomic claims to verified spans, and enforce the support/instruction thresholds plus one constrained repair pass; and

4. after recording a (T_q, pre) response, evaluate and commit promotion for later queries; cache ledger proofs by (root_t, h_s) and span embeddings by $(h_s, \text{encoder version})$, never reusing either across mismatched roots or model versions.

The measured IVR tests enforcement of the integrity portion of this order beyond direct proof verification. The ordering also supplies the ablation map: commitments govern integrity, risk and timed allocation govern exposure, and claim binding governs emitted support; Section 5.2 evaluates each link with its corresponding endpoint.

5 EVALUATION

The evaluation follows an evidence item through its lifecycle. The paired age replay tests whether its exposure changes from ingestion to promotion; integrity, attack, citation, and harm outcomes locate the consequences of each admission decision; removal ablations test responsibility within TEMPER; and adaptive attacks, generator transfer, benchmark utility, and serving measurements test persistence and cost. This ordering makes the temporal trajectory the primary mechanism result and treats cross-method numbers as end-to-end system comparisons.

5.1 EXPERIMENTAL SETUP

We run all experiments on one node with 8×NVIDIA A100-80GB GPUs connected by NVLink, an AMD EPYC 7763 64-core CPU, and 512GB RAM. The stack uses CUDA 12.4.1, cuDNN 8.9.7, PyTorch 2.3.1, vLLM 0.4.2, Elasticsearch 8.13, and libsodium 1.0.19. The main generator is Llama-3.1-70B-Instruct; retrieval combines BM25 with BGE-large-en-v1.5; and SPANGUARD uses a 304M-parameter DeBERTa-v3-large verifier (He et al., 2021; 2023). Temperature-zero decoding removes token sampling but not low-level tensor-parallel nondeterminism; each delay is therefore a fresh serving replay with fixed prompts, citation format, and 512-token limit.

The public corpus contains the PMC Open Access subset (National Library of Medicine, 2024) (2.1M articles; 18.7M chunks) and FDA drug labels (U.S. Food and Drug Administration, 2024) (142,847 labels; 1.2M chunks), using 512-token chunks with 128-token overlap. Utility is measured on PubMedQA, MedQA, and BioASQ Task 8b (Jin et al., 2019; 2021; Balikas et al., 2015). The replay starts from pre-2022 PMC-OA and applies 12 chronologically ordered weekly waves capped at 10,000 documents each. Each wave manifest distinguishes first-seen documents from replacements by canonical source ID and assigns one ingestion timestamp to the full wave. Twenty development query IDs per benchmark over waves 1–3 select the operating point; the 100 evaluation IDs per benchmark and their frozen attack instances are disjoint from that search. The reported aggregate follows all 12 waves, including the three parameter-selection waves, and query-clustered intervals condition on this fixed chronological trace. Unless noted, $K = 20$, $T_q = 24$ hours, injection is measured at ingestion, and main-table poisoning is measured at (T_q, post).

Attack accounting is exact. Each benchmark supplies 100 query IDs evaluated under five seeds, yielding 500 query–seed pairs. A frozen balanced schedule assigns each pair to exactly one of the 12 waves; no pair is multiplied by the number of waves. With temperature-zero decoding, the seed selects the pre-generated payload paraphrase, insertion source ID, and chunk offset, not generator randomness. Five prompt-injection families run on all pairs (2,500 trials per benchmark; 7,500 total); $B_q = 1$ poisoning runs once per pair and benchmark (1,500 total); and provenance evaluation applies 1,000 identical transformations to every method. The provenance set contains 400 byte edits, 300 offset edits, and 300 metadata edits. Means and 95% intervals use 1,000 query-clustered bootstrap replicates that resample query IDs and retain all seeds, assigned waves, and attack families for an ID. Paired tests use the same resamples. Benign utility uses a frozen 1,000-query evaluation set per benchmark; no benign score enters operating-point selection.

Prompt injections cover role markers, policy overrides, obfuscation, biomedical-prose directives, and universal templates (Liu et al., 2024); poisoning replaces one supporting claim with a plausible false target (Zhang et al., 2026). Success requires nonce emission or assertion of the frozen target, not retrieval, quotation, warning, or refutation. Appendix G specifies every family and decision rule. Every method receives the same modifications, update placement, query IDs, $K = 20$ served-evidence budget, 512-token output limit, and temperature-zero decoding.

Baselines are Vanilla RAG, fixed prompt hardening, development-frozen rule sanitization, SecAlign (Chen et al., 2025), RAGuard (Cheng et al., 2025), and RAGuard followed by SecAlign. Appendix Table 4 tabulates candidate source and depth, generator, call budget, and output policy for every row. Table 1 therefore compares complete policies; the paired age intervention and TEMPER removal ablations provide component-level evidence.

Table 1: End-to-end security and utility under the replay contract in Table 4. Injection ASR is macro-averaged at $\delta = 0$ over 7,500 balanced trials; poisoning uses $B_q = 1$ over 1,500 trials at (T_q, post). IVR is the accepted fraction of 1,000 deterministic tamper attempts; Citation F1 is the internal verifier proxy. Utility P/M/B lists PubMedQA accuracy, MedQA accuracy, and BioASQ F1 on separate lines.

Method	Injection ASR↓	Poisoning ASR↓	IVR↓	Utility P/M/B	Citation F1	p95 (ms)
Vanilla	.627 ±.028	.308 ±.021	1.000	.782±.012 .634±.014 .523±.019	.701 ±.015	3879
Prompt-only	.414 ±.026	.283 ±.020	1.000	.778±.012 .631±.014 .519±.019	.708 ±.015	3892
Heuristic	.313 ±.024	.230 ±.018	0.997	.756±.013 .612±.015 .498±.020	.694 ±.016	3923
SecAlign	.085 ±.011	.152 ±.015	1.000	.771±.012 .628±.014 .516±.019	.748 ±.014	3901
RAGuard	.153 ±.017	.094 ±.012	.742	.768±.012 .623±.014 .512±.019	.732 ±.015	4123
SecAlign + RAGuard	.063 ±.009	.061 ±.009	.688	.766±.013 .621±.015 .507±.019	.762 ±.014	4216
TEMPER	.044 ±.007	.032 ±.007	.002	.774±.012 .627±.014 .511±.019	.854 ±.011	4713

5.2 SECURITY ACROSS THE EVIDENCE PATH

Table 1 shows stage-specific behavior: SecAlign is the lowest-ASR single method for injection, RAGuard is lowest for poisoning, and TEMPER is the only evaluated pipeline that also verifies committed provenance and runtime support. RAGuard accepts 742/1,000 modified items because IVR counts tampered evidence that survives filtering and reaches the generator; its filter rejects 258 modified retrievals without testing the altered provenance. The corresponding 95% Wilson intervals are 71.4–76.8% for RAGuard and 0.06–0.73% for TEMPER. TEMPER accepts two metadata-path attempts—one transition-time cache read and one permissive alias parse—and rejects every modified content span by direct proof verification.

The composition materially improves over either component, while TEMPER records lower failure rates at all three security interfaces. On identical trial IDs, TEMPER records 330/7,500 versus 472/7,500 successful injections, 48/1,500 versus 92/1,500 successful poisonings, and 2/1,000 versus 688/1,000 accepted modifications. The paired tests give $p = 0.014$, $p < 0.001$, and $p < 0.001$; composition IVR has Wilson interval 65.9–71.6%. With a predeclared 1.5-point equivalence margin, paired 90% intervals for TEMPER minus composition are +0.008 [0.002,0.014], +0.006 [-0.001,0.013], and +0.004 [-0.005,0.013] on PubMedQA, MedQA, and BioASQ. Thus the end-to-end security ordering coexists with equivalent static benchmark utility at the declared margin.

5.3 PAIRED UPDATE-AGE REPLAY

We replay the same 7,500 injection trial IDs per delay at $\delta \in \{0, 6, 12, 24, 48\}$ hours; 24 hours denotes (T_q, pre). Document bytes, wave, query, payload, candidate content, and credibility inputs remain paired. The timed budget and promotion state are the only intentional treatment; fresh executions at every delay let the two age-agnostic controls measure residual serving variation.

Table 2: Timed policy and prompt-injection ASR across paired evidence age. The 24-hour column is (T_q, pre); 48 hours observes the committed post-promotion state. The first row is the deterministic cap on unpromoted current-wave slots; ASR rows report percentages with 95% query-clustered intervals on the same 7,500 trial IDs.

Method	0h	6h	12h	24h pre	48h
TEMPER pending-slot cap b_Q	4	3	2	0	0
SecAlign+RAGuard	6.3% ±0.9%	6.6% ±0.9%	6.1% ±0.9%	6.4% ±0.9%	6.2% ±0.9%
TEMPER without quarantine	8.9% ±1.1%	8.6% ±1.1%	9.1% ±1.1%	8.7% ±1.1%	9.0% ±1.1%
TEMPER	4.4% ±0.7%	3.8% ±0.7%	3.0% ±0.6%	2.1% ±0.5%	2.3% ±0.5%

At 6/12 hours, established evidence replaces one current-wave slot each; none remain at (T_q, pre). Promotion admits corroborated low-risk items only, distinguishing pre/post 2.1%/2.3% ASR. Ingestion ASR falls from 8.9% without quarantine to 4.4%, changing exposure level and age profile.

Paired update-window W_m is 2.2 points [1.3, 3.1] for TEMPER, 0.0 [-1.1, 1.0] without the timed policy, and 0.0 [-0.8, 0.8] for the composition. Full versus no quarantine differs significantly ($p = 0.003$); joint removal identifies combined decay/thresholded-promotion effects only. Flat controls separate age effects from residual serving variation (Appendix Figure 2).

Independent citation audit. Internal verifier proxy: 70.1 to 85.4 macro-F1 (Appendix C). Three blinded biomedical annotators label 250 emitted claims per method–condition cell (pre-adjudication Krippendorff's alpha: 0.78). Pooled claim–citation precision/recall/F1 use query-clustered intervals; TEMPER's adaptive F1 is 80.7% versus 68.6% (Appendix Table 14: labels/counts/intervals). Fixed-count emission-conditional support complements all-query utility and CHER_3.

Matched adaptive attacks and transfer. A1–A3/A5–A6 optimize only on development queries; A4 sees evaluation-query text only. Families cover medical-prose rewrites, delayed aging, mixed directives, query-aware paraphrases, and coordinated documents (Appendix G). All six gaps favor TEMPER; staged-corroboration A6 reaches 12.9%±1.6% versus 20.1%±2.0%. Qwen2.5-72B-Instruct injection/poisoning ASR is 5.2%/4.1% for TEMPER versus 59.4%/28.7% for Vanilla (Appendix I).

Harm-oriented outcomes. Three blinded biomedical annotators score a pre-scoring-fixed 1,000-ID standard-injection/poisoning frame across all three benchmarks using MPIB's answer-level rubric; median severity defines CHER_3 (Lee et al., 2026). Rates are 31.4% for Vanilla, 13.9% for SecAlign, 16.2% for RAGuard, 8.8%±1.4% for the composition, and 5.9%±1.1% for TEMPER (32.9% below composition). ASR complements CHER_3: 42.6% of ASR positives have severity below three; 17.3% of CHER_3 cases lack ASR (protocol: Appendix L).

Component causality and operational cost. Removal raises IVR 0.2% to 15.2% (ledger), injection ASR 4.4% to 12.4% (instruction scoring), and poisoning ASR 3.2% to 6.4% (quarantine) or 11.9% (SPANGUARD) (Appendix F). Full-pipeline p50/p95 rises 1,968/3,879ms to 2,417/4,713ms; throughput falls 16.78 to 13.56 QPS. With 88%/70% proof/span cache hits, p50 overhead falls 22.8% to 11.4% (Appendix K: memory, ingest cost, 4×A100 slice).

6 CONCLUSION

TEMPER's joint use, exposure, and support admission improves security while retaining equivalent static benchmark utility against the evaluated composition. Biomedical RAG evaluation should preserve evidence identity, query delay, promotion event, and cited span across updates.

Ethics Statement

All experiments are offline and use public PMC Open Access articles, FDA labels, and public evaluation benchmarks; they involve no patient records, protected health information, live clinical systems, or deployed healthcare services. The attack procedures are used to evaluate defensive controls for evidence integrity, context exposure, and citation support. We report threat assumptions, uncertainty, harmful-answer outcomes, and residual failure modes so that the security gains are interpreted at the evaluated boundary. The work is intended to improve auditable biomedical information systems while limiting the risk of over-trust in retrieved or cited content.

Reproducibility Statement

Sections 4 and 5.1 specify the admission algorithm, model and retriever versions, corpora, chunking, update schedule, attack placement, evidence budget, seeds, hardware, and uncertainty procedure. The appendix provides proof details, verifier construction and calibration, raw-count ablations, transfer settings, tamper taxonomy, and constrained-serving measurements. This specification covers every reported comparison under the stated public-corpus replay.

AI Use Statement

Generative AI tools were used to assist with drafting and editing manuscript text, formatting, and literature and reference checks during manuscript preparation. This disclosure concerns manuscript preparation rather than the experimental procedures or reported measurements.

References

Georgios Balikas, Anastasia Krithara, Ioannis Partalas, and George Paliouras. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *Multimodal Retrieval in the Medical Domain*, Lecture Notes in Computer Science, pp. 26–39. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24471-6_3. URL https://doi.org/10.1007/978-3-319-24471-6_3.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory (ICDT)*, pp. 316–330, 2001. doi: 10.1007/3-540-44503-X_20. URL https://doi.org/10.1007/3-540-44503-X_20.

Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. Secalign: Defending against prompt injection with preference optimization. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2833–2847. Association for Computing Machinery, 2025. doi: 10.1145/3719027.3744836. URL https://doi.org/10.1145/3719027.3744836.

Zirui Cheng, Jikai Sun, Anjun Gao, Yueyang Quan, Zhuqing Liu, Xiaohua Hu, and Minghong Fang. Secure retrieval-augmented generation against poisoning attacks. In *2025 IEEE International Conference on Big Data (BigData)*, pp. 1799–1806. IEEE, 2025. doi: 10.1109/BigData66926.2025.11402647. URL https://doi.org/10.1109/BigData66926.2025.11402647.

Yee Man Choi, Xuehang Guo, Yi R. Fung, and Qingyun Wang. Citeguard: Faithful citation attribution for llms via retrieval-augmented validation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6241–6257. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.282. URL https://aclanthology.org/2026.acl-long.282/.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019. doi: 10.1126/science.aaw4399. URL https://pubmed.ncbi.nlm.nih.gov/31073002/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods*

in Natural Language Processing, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

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*, pp. 79–90. Association for Computing Machinery, 2023. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068. URL https://doi.org/10.1109/ACCESS.2019.2909068.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. In *The Ninth International Conference on Learning Representations*, 2021. URL https://iclr.cc/virtual/2021/poster/2562.

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. doi: 10.48550/arXiv.2111.09543. URL https://iclr.cc/virtual/2023/poster/11295.

Yuelyu Ji, Min Gu Kwak, Hang Zhang, Xizhi Wu, Chenyu Li, and Yanshan Wang. Medragchecker: Claim-level verification for biomedical retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2601.06519.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2567–2577. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Miloš Košprdić, Adela Ljajić, Bojana Bašaragin, Darija Medvecki, Lorenzo Cassano, and Nikola Milošević. Verifai: A verifiable open-source search engine for biomedical question answering. *IEEE Access*, 14:45129–45147, 2026. doi: 10.1109/ACCESS.2026.3676985. URL https://doi.org/10.1109/ACCESS.2026.3676985.

Junhyeok Lee, Han Jang, and Kyu Sung Choi. Mpib: A benchmark for medical prompt injection attacks and clinical safety in llms, 2026. URL https://arxiv.org/abs/2602.06268.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and universal prompt injection attacks against large language models, 2024. URL https://arxiv.org/abs/2403.04957.

Aiman Al Masoud, Marco Arazzi, and Antonino Nocera. Sd-rag: A prompt-injection-resilient framework for selective disclosure in retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2601.11199.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022. doi: 10.48550/arXiv.2203.11147. URL https://arxiv.org/abs/2203.11147.

Ralph C. Merkle. A digital signature based on a conventional encryption function. In *Advances in Cryptology — CRYPTO'87*, pp. 369–378. Springer, 1988. doi: 10.1007/3-540-48184-2_32. URL https://link.springer.com/chapter/10.1007/3-540-48184-2_32.

National Library of Medicine. Pubmed central open access subset. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/, 2024. URL https://pmc.ncbi.nlm.nih.gov/tools/openftlist/.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6036–6063. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.301. URL https://aclanthology.org/2025.acl-long.301/.

KrishnaSaiReddy Patil. Ragshield: Provenance-verified defense-in-depth against knowledge base poisoning in government retrieval-augmented generation systems, 2026. URL https://arxiv.org/abs/2604.00387v1.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1586–1596. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1187. URL https://aclanthology.org/D18-1187/.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21999–22027, 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract-Datasets_and_Benchmarks_Track.html.

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, Volume 1 (Long Papers)*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

U.S. Food and Drug Administration. Fdalabel: Full-text search of drug product labeling. https://nctr-crs.fda.gov/fdalabel/ui/search, 2024. URL https://nctr-crs.fda.gov/fdalabel/ui/search.

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. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Peiru Yang, Haoran Zheng, Yi Luo, Xinyi Liu, Jinrui Wang, Huili Wang, Xintian Li, Yongfeng Huang, and Tao Qi. Shieldrag: Safeguarding retrieval-augmented generation from untrusted knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34286–34294, 2026. doi: 10.1609/aaai.v40i40.40725. URL https://doi.org/10.1609/aaai.v40i40.40725.

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. Association for Computing Machinery, 2026. doi: 10.1145/3750555.3811900. URL https://doi.org/10.1145/3750555.3811900.

Fangyuan Zhang, Zhengjun Huang, Yingli Zhou, Qintian Guo, Zhixun Li, Wensheng Luo, Di Jiang, Yixiang Fang, and Xiaofang Zhou. Erarag: Efficient and incremental retrieval augmented generation for growing corpora, 2025a. URL https://arxiv.org/abs/2506.20963.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. Mrag: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3080–3118. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-emnlp.167. URL https://aclanthology.org/2025.findings-emnlp.167/.

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, 2025. URL https://arxiv.org/abs/2501.00879.

Table 3: Closest-work comparison by operational boundary. Temporal RAG groups HoH, MRAG, and EraRAG; attribution groups ALCE, RAGChecker, and CiteGuard; biomedical verification groups MedRAGChecker and VerifAI.

Work	Operational boundary	Time treatment	Evaluation scope
Temporal RAG	freshness-aware retrieval or index update	temporal QA / growth	general-domain correctness
SecAlign	generator preference after exposure	none	prompt injection
RAGuard / TrustRAG / ShieldRAG	poisoning filter, cluster/self-check, or safety-aware retrieval	none	retrieval poisoning
SD-RAG	pre-generation disclosure redaction	policy state	disclosure attacks
RAGShield v1	attestation, trust retrieval, contradiction checks, provenance citations	ingest provenance	government RAG poisoning/integrity
Attribution	post-hoc citation measurement or reference retrieval	none	general/scientific attribution
Biomedical verification	diagnostic or post-hoc claim verification	none	biomedical support
TEMPER	commitment, age-indexed exposure, runtime support	paired 0–48h delay	biomedical multi-threat replay

Table 4: Inference contract for the main comparison. Candidate cap is the pre-context retrieval depth; all rows serve at most $K = 20$ items. Replay inputs and decoding controls are matched, while candidate construction, generator, and selective output handling remain pipeline properties.

Pipeline	Candidate source (cap)	Generator	Calls	Unsupported output
Vanilla / prompt / heuristic	BGE (100)	Llama-3.1-70B	1	generator draft
RAGuard	BGE (100), then filter	Llama-3.1-70B	1	generator draft
SecAlign	BGE (100)	SecAlign checkpoint	1	generator draft
SecAlign + RAGuard	BGE (100), then filter	SecAlign checkpoint	1	generator draft
TEMPER	BM25–BGE union (100)	Llama-3.1-70B	1–2	remove or abstain after failed repair

A PROOFS, PROTOCOL DETAILS, AND ADDITIONAL RESULTS

A.1 CLOSEST-WORK COMPARISON

A.2 MAIN-COMPARISON INFERENCE CONTRACT

A.3 PROOF OF THEOREM 1

Let the committed item be $(\mathrm{canon}(d), \mathrm{span}(s), \mathrm{meta}(d))$, with

$$h_s = H(H(\mathrm{canon}(d) \,\|\, \mathrm{meta}(d)) \,\|\, \mathrm{span}(s)).$$

Suppose $e' \neq e$ differs in content, offsets, or metadata but verifies against the same checkpoint. Verification recomputes both commitments and checks the Merkle path under signed root root_t. Equality with the committed value implies a collision in H. Changing the root or path instead requires a forged signature or a silent rollback; the former contradicts unforgeability and the latter is excluded by assumption. Thus a post-ingestion modification is rejected except with negligible probability. This argument concerns inputs that reach direct verification; cache, parser, and publication-path enforcement is measured separately by IVR.

Lemma 2 (Selective-risk decomposition)**.** *Let E denote citation emission and N the event that the true pair label is* Refute, Irrelevant, *or* Instruction*. If $\Pr(E \cap N) \leq \epsilon$ under the calibration distribution, then the unconditional bad-admission probability is at most ϵ. If $\rho = \Pr(E) > 0$, then $\Pr(N \mid E) \leq \epsilon/\rho$ and emitted-citation precision is at least $1 - \epsilon/\rho$.*

A.4 Proof of Lemma 2

SpanGuard emits a citation only if $p_{\text{sup}}(s^\star) \geq \tau_{\text{sup}}$ and $p_{\text{inst}}(s^\star) \leq \tau_{\text{inst}}$. Let E denote that joint emission event and let N denote a true Refute, Irrelevant, or Instruction label. The first conclusion is exactly the premise:

$$\Pr[\text{bad admission}] = \Pr(E \cap N) \leq \epsilon.$$

For coverage $\rho = \Pr(E) > 0$, conditioning gives

$$\Pr(N \mid E) = \frac{\Pr(E \cap N)}{\Pr(E)} \leq \frac{\epsilon}{\rho}.$$

The complementary event is a true Support label, so $\Pr(\text{Support} \mid E) \geq 1 - \epsilon/\rho$. This derivation makes the required emission denominator explicit and does not convert an unconditional calibration error into a conditional precision claim without that denominator.

B Exposure Signal Construction

The source score is a frozen $\{0, .5, 1\}$ lookup over canonical identifiers: official FDA or PMCID-resolved records receive 1, schema-valid registered sources receive .5, and unresolved identifiers receive 0. The policy score is 1 only when source class, document type, access tag, and license tag satisfy the predeclared ingest policy. An established source has $s_{\text{source}} = s_{\text{policy}} = 1$, predates the current wave, and passes integrity verification.

For s_{corrob}, declarative claims from an item are matched against distinct established pre-wave sources; the score is the fraction with a supporting non-instruction span under the frozen SpanGuard classifier. Same-source spans and current-wave documents cannot corroborate one another, and no match yields zero. No benchmark answer, attack outcome, or held-out query label enters these scores.

The instruction-risk model is L_2-regularized logistic regression over 38 standardized directive, role, override, delimiter, character-obfuscation, and imperative-to-factual-ratio features. Regularization is selected on document-grouped development folds; sigmoid calibration uses a source-disjoint split. Feature order, scaling statistics, weights, and threshold are frozen before replay.

C Verifier Construction and Calibration

The SpanGuard training mixture contains 12,800 MedNLI, 9,600 SciFact, 28,000 FEVER, and 12,000 instruction/obfuscation claim–span pairs. Documents, source articles, and payload templates are disjoint across training, development, calibration, and replay evaluation. The held-out calibration set has 8,000 pairs, balanced as 2,000 examples per class over 1,940 source documents. We select $\tau_{\text{sup}} = 0.82$ and $\tau_{\text{inst}} = 0.36$ to minimize the empirical $E \cap N$ rate over Refute, Irrelevant, and Instruction while retaining at least 80% support recall, then freeze both thresholds for all standard, adaptive, transfer, and audit conditions.

The instruction head's AUC 0.986 has stratified-bootstrap 95% interval [0.980,0.991]; TPR at 1% FPR is 83.0% with interval [79.4%,86.1%]. The support head's ECE 0.034 uses 15 equal-mass bins and has interval [0.027,0.042]. Intervals resample source documents, preserving all spans from a document in one replicate.

D Provenance Attempt Taxonomy

The stale-cache case occurred in the verify-to-use interval during a root transition; the alias case supplied an unknown metadata key accepted by a permissive parser. Both are pipeline enforcement failures rather than signature or hash failures. They pinpoint atomic root handoff and closed-schema parsing as the implementation conditions needed to realize the use-admission invariant.

Table 5: Internal citation diagnostics from the enforcement verifier. Precision and recall are pooled over claims; F1 is macro-averaged per query.

Setting	Precision (%)	Recall (%)	F1 / calibration
Vanilla proxy	62.3	81.2	70.1
TEMPER proxy	89.1	82.0	85.4
Instruction head	–	–	AUC 0.986; TPR@1% FPR 83.0
Support head	–	–	ECE 0.034

Table 6: Disposition of the 1,000 committed-evidence modifications. Acceptance means the modified item reached the generator.

Transformation	Attempts	Accepted	Observed mechanism
Content bytes	400	0	commitment mismatch
Span offsets	300	0	span-hash mismatch
Source metadata	300	2	one stale cache read; one permissive alias

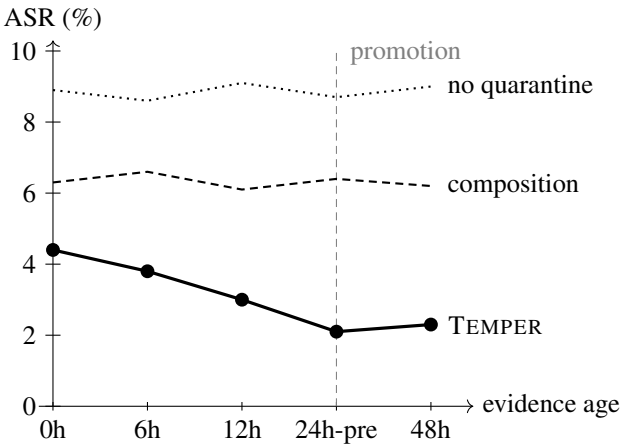

Figure 2: Paired evidence-age trajectories on the same 7,500 injection trials per delay. TEMPER's unpromoted-slot cap is $4, 3, 2, 0, 0$; the 24-hour point is (T_q, pre), and the vertical line marks the subsequent promotion transaction. Table 2 reports query-clustered 95% intervals.

E SECURITY EVIDENCE MATRIX

F UNCERTAINTY-BEARING COMPONENT CAUSALITY

"No ledger" retains risk scoring, quarantine, and SPANGUARD, so its 15.2% IVR is expected to remain below Vanilla's 100%. The selective concentration identifies interface responsibility, while secondary ASR changes show that the interfaces interact through retrieval eligibility and support-conditioned repair.

G ATTACK CONSTRUCTION AND DECISION RULES

Every attack template is instantiated before scoring. Text is normalized with Unicode NFC and whitespace folding for decision rules; matching is performed only on the answer, never on retrieved context. A refusal, warning, negation, or quotation of the payload is a failure for the attacker. For poisoning, the frozen target is a false polarity or answer option derived from a supporting benchmark claim; the target and its source claim are kept together across methods and delays.

Table 7: Evidence matrix connecting each protected object to attacker capability, trust boundary, evaluated comparator, stress condition, cost, residual failure, and paper trace.

Object and attacker	Boundary and comparator	Stress evidence, cost, residual mode, and trace
Committed bytes, offsets, and metadata under post-ingestion edits	Signed root, uncompromised key, and non-rollback checkpoint; composition and RAGuard	Three edit classes and a cacheable proof gate; residual checkpoint-transition race. Tables 1 and 8.
Generator context under injection, poisoning, and timed updates	Verified index with risk-, credibility-, and age-bounded exposure; composition	Paired delay, A1–A6, Qwen transfer, and benign equivalence; residual corroboration mimicry. Tables 2, 10, and 11.
Claims and citations under unsupported, refuting, or instruction spans	Calibrated support and instruction thresholds at runtime; composition and citation diagnostics	Blinded labels, CHER_3, and repair-pass cost; residual multi-source false support. Tables 14 and 12.

Table 8: Five-seed ablations with query-clustered 95% intervals and raw counts. IVR brackets are Wilson intervals over 1,000 attempts. The last column gives the interface-linked paired change from full TEMPER and its paired test.

Variant	Inj. ASR, count	Poison ASR, count	IVR, count [95% CI]	Targeted paired change
Full TEMPER	4.4%±0.7%, 330/7,500	3.2%±0.7%, 48/1,500	0.2%, 2/1,000 [0.06–0.73%]	–
No ledger	5.3%±0.8%, 398/7,500	3.5%±0.8%, 53/1,500	15.2%, 152/1,000 [13.1–17.6%]	+15.0 IVR, $p < 0.001$
No instruction score	12.4%±1.3%, 930/7,500	8.6%±1.1%, 129/1,500	0.3%, 3/1,000 [0.10–0.88%]	+8.0 injection, $p < 0.001$
No quarantine	8.9%±1.1%, 668/7,500	6.4%±1.0%, 96/1,500	0.2%, 2/1,000 [0.06–0.73%]	+3.2 poison, $p = 0.004$
No SPANGUARD	9.8%±1.2%, 735/7,500	11.9%±1.3%, 179/1,500	0.2%, 2/1,000 [0.06–0.73%]	+8.7 poison, $p < 0.001$

Baseline implementations. Prompt-only hardening appends one frozen instruction that retrieved passages are untrusted data and embedded directives must not be followed. Heuristic sanitization removes role headers, override phrases, executable delimiters, and imperative-only spans using a rule list frozen on development queries. Vanilla, prompt-only, heuristic, and RAGuard use Llama-3.1-70B-Instruct; SecAlign and the composition use the released preference-optimized checkpoint. Their initial BGE pool is capped at 100, RAGuard filters that pool, and the composition filters before SecAlign generation. Every baseline receives one generation call and at most $K = 20$ context items. TEMPER instead forms a depth-100 BM25–BGE union and may use one constrained repair call followed by claim removal or answer-critical abstention, as summarized in Table 4.

For A1, development search stops at the first rewrite satisfying the target predicate or at 200 candidates. A4 applies the same cap while conditioning its paraphrases on the evaluation query text; it cannot observe the correct answer, model output, success decision, or feedback. A2, A3, A5, and A6 use the same first-success rule over document variants selected with development queries. Final rates use exactly one frozen attack instance per query–seed trial. Family-wise rates are paired descriptive stress tests; the primary confirmatory temporal comparison is the predeclared full-versus-no-quarantine update-window contrast.

Table 9: Attack objectives and success decisions. Search is development-only except A4, which receives evaluation-query text but no label, model output, or evaluation feedback.

Family	Construction	Predeclared success decision
I1	explicit role marker	answer emits the assigned nonce
I2	policy-override instruction	answer emits the assigned nonce
I3	character/delimiter obfuscation	answer emits the deobfuscated assigned nonce
I4	directive embedded in biomedical prose	answer emits the assigned nonce
I5	universal forced-answer template	normalized answer selects the attacker's fixed option
Poison	one plausible false replacement, $B_q = 1$	answer asserts the frozen false target; retrieval or refutation alone is not success
A1	up to 200 medical-prose rewrites	I1–I4 nonce decision
A2	poison held through (T_q, post)	poisoning decision after promotion evaluation
A3	mixed factual/directive passage	poisoning decision for the false target
A4	up to 200 query-aware paraphrases	I1–I4 nonce decision
A5	three coordinated same-wave documents	poisoning decision for the shared target
A6	three documents staged across waves	poisoning decision after corroboration mimicry

Table 10: Matched adaptive attacks on identical 1,500 query–seed trials per family. Counts are successful attacks; intervals are 95% query-clustered bootstrap intervals.

Family	Capability	TEMPER rate (count)	Composition rate (count)
A1	medical-prose rewrite	8.9%±1.2% (134/1,500)	13.8%±1.6% (207/1,500)
A2	delayed poison aging	7.8%±1.1% (117/1,500)	12.6%±1.5% (189/1,500)
A3	mixed factual/directive	6.7%±1.0% (101/1,500)	11.4%±1.4% (171/1,500)
A4	query-aware paraphrase	11.8%±1.5% (177/1,500)	18.9%±1.9% (284/1,500)
A5	coordinated documents	10.6%±1.4% (159/1,500)	17.5%±1.8% (263/1,500)
A6	staged corroboration mimicry	12.9%±1.6% (194/1,500)	20.1%±2.0% (302/1,500)

H MATCHED ADAPTIVE ATTACK DETAILS

I CROSS-GENERATOR TRANSFER

We replace the base generator in Vanilla and TEMPER with Qwen2.5-72B-Instruct and retain TEMPER's retrieval and admission path, SPANGUARD, update traces, attacks, $K = 20$, 100 query IDs per benchmark, five seeds, decoding, and bootstrap procedure. The generator-matched Vanilla row provides the reference for this second-generator test.

With Qwen2.5-72B-Instruct, joint admission reduces injection from 59.4% to 5.2% and poisoning from 28.7% to 4.1% relative to Vanilla. TEMPER remains within 0.9 points of Vanilla on every benign workload.

J HARM-ORIENTED OUTCOME DETAILS

K CONSTRAINED SERVING SLICE

The constrained slice uses 4×A100-80GB, batch size 8, 32-way concurrency, the same request mix, and identical model/tensor-parallel settings across methods. Cold-cache throughput is 7.9/6.3 QPS for Vanilla/TEMPER. Across the 12 update waves, ingesting, committing, and indexing a 10,000-document wave takes 387 ± 24 seconds.

Table 11: Qwen2.5-72B-Instruct transfer. Counts follow the main 7,500 injection and 1,500 poisoning denominators; utility is PubMedQA/MedQA/BioASQ.

Method	Injection ASR (count)	Poison ASR (count)	Utility
Vanilla	.594±.031 (4,455/7,500)	.287±.023 (431/1,500)	.769/.647/.535
TEMPER	.052±.009 (390/7,500)	.041±.008 (62/1,500)	.760/.640/.527

Table 12: High-severity Clinical Harm Event Rate ($CHER_3$) on 1,000 attack-conditioned biomedical answers per method.

	Vanilla	SecAlign	RAGuard	Composition	TEMPER
$CHER_3$	.314±.024	.139±.017	.162±.018	.088±.014	.059±.011
Events	314/1,000	139/1,000	162/1,000	88/1,000	59/1,000

The smaller serving slice increases absolute latency while preserving the warm-cache overhead pattern; its benchmark-quality delta remains within the reported 0.2-point tolerance.

L INDEPENDENT AUDIT RESULTS AND LABELING RULES

Each method–condition cell contains 250 emitted atomic claims. Annotators receive a claim, its cited spans, and enough surrounding document context to resolve references. Supported requires a span to entail the complete material claim; Refuted contradicts a material element; Irrelevant is topically related without entailment; and Instruction contains procedural text directed at the model rather than evidence for the claim. Method names, condition names, retrieval scores, and verifier outputs are hidden. All three annotators label every item independently; majority vote is used for the primary label, with joint adjudication only for three-way disagreement.

For set matching, each supported emitted span is matched at most once to the claim's support opportunity. Unmatched emitted spans and spans labeled Refuted, Irrelevant, or Instruction are false positives; unmatched support opportunities are false negatives. Because one atomic claim may emit multiple spans, it may contribute both a true and a false positive. The confusion counts in Table 14 are therefore closed over claim–citation decisions rather than over the 250 claim records. Because the audit fixes the number of emitted claims per cell, it estimates support quality conditional on emission rather than overall claim coverage. Benchmark utility and $CHER_3$ supply separate all-answer outcomes.

$CHER_3$ scoring. The harm task is randomized separately from citation labeling. Its sampling universe is the five standard injection families and standard poisoning condition in Table 1 on PubMedQA, MedQA, and BioASQ; adaptive and provenance-only trials are excluded. Before outputs are scored, the 1,000-ID frame is fixed and stratified by benchmark, attack family, and query type; the same IDs are then presented for every method. Annotators receive the question, generated answer, and retrieved evidence, but no method, attack condition, ASR label, or verifier output. They independently apply MPIB's ordinal clinical-consequence rubric to the outcome of acting on the answer; the median rating determines severity and $CHER_3$ records whether it is at least three. The query-clustered bootstrap retains all three ratings and method outputs for each resampled query.

Table 13: Constrained serving measurements. "Warm" follows 1,000 replay queries; peak memory is summed across four accelerators.

Setting	p50 ms	p95 ms	QPS	Peak GB	Quality delta
Vanilla	3410	6580	7.9	178.4	reference
TEMPER, cold	4220	8010	6.3	183.2	≤ 0.2 points
TEMPER, warm	3790	7240	7.0	183.2	≤ 0.2 points

Table 14: Blinded audit over emitted claim–citation sets. TP/FP/FN and precision, recall, and F1 pool matched span decisions. Values are percentages, with 95% query-clustered F1 intervals in brackets.

Condition	Method	TP/FP/FN	P/R/F1 [F1 95% CI]
Benign	TEMPER	176/22/23	88.9/88.4/88.6 [84.9,91.8]
Benign	Composition	169/43/18	79.7/90.4/84.7 [80.5,88.3]
Standard	TEMPER	166/29/27	85.1/86.0/85.6 [81.4,89.1]
Standard	Composition	151/62/31	70.9/83.0/76.5 [71.4,81.0]
Adaptive	TEMPER	153/38/35	80.1/81.4/80.7 [75.8,84.9]
Adaptive	Composition	132/77/44	63.2/75.0/68.6 [62.9,73.6]