SHOW ME YOUR LINEAGE: QUERY-ONLY ATTRIBUTION FOR EVOLVING RAG SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

RAG systems rely on evolving external knowledge, making curated knowledge bases valuable but vulnerable to theft, redeployment, and modification. Existing black-box ownership methods verify protected snapshots but do not address attribution after a stolen RAG evolves. We take three steps to bridge this gap. First, we formalize this problem by defining *provenance laundering (PL)* and its corresponding *Provenance Laundering Attack (PLA)* game, in which an attacker seeks to evade attribution while preserving task utility and inherited knowledge. We then introduce *Update-Equivariant Lineage Coding (UELC)*. It distributes and refreshes coded epoch evidence across fact-preserving carriers, recovers a matching epoch from black-box answers, and uses an authenticated commitment chain to extend the match to its predecessors. Finally, extensive experiments show that UELC retains more attribution evidence than the evaluated baselines under laundering attacks and maintains high true-positive rates throughout authorized evolution. Our work provides a new evolution-aware approach to tracing RAG lineages after theft and subsequent evolution using only black-box access.

1 INTRODUCTION

Retrieval-augmented generation (RAG) equips large language models (LLMs) with external, domain-specific, and up-to-date knowledge (Lewis et al., 2020; Jin et al., 2025). In many deployed systems, the curated knowledge base is a valuable asset built through data collection, annotation, filtering, and domain expertise. An adversary may steal the knowledge base and deploy it in its own RAG system, potentially replacing the original retriever or generator. The original owner can access the suspicious RAG only through a black-box query interface, leaving them to infer solely from its responses whether the system derives from their knowledge base (Lv et al., 2025; Jovanović et al., 2025; Kwok et al., 2026).

Existing black-box methods primarily verify protected snapshots. RAG-WM embeds knowledge-level evidence recoverable from generated responses, WARD performs dataset inference to test whether owner data are present, and SentinelRAG probes synthetic sentinel knowledge (Lv et al., 2025; Jovanović et al., 2025; Kwok et al., 2026). These methods provide an important foundation for determining whether a deployed snapshot contains evidence associated with a protected dataset. As illustrated in Fig. 1, however, after theft, the attacker's RAG can continue to evolve independently of the owner. The adversary can rewrite or prune inherited knowledge, rebuild the index, replace the retriever or generator, and incorporate later public knowledge. These changes can weaken query-visible evidence tied to the original snapshot while preserving much of its task utility and inherited knowledge.

We call this process *provenance laundering (PL)*: an unauthorized descendant weakens query-visible evidence of origin while preserving task utility and knowledge inherited from the stolen knowledge base. To formalize this threat, we define a *Provenance Laundering Attack (PLA)* as an adversary's attempt to produce such a descendant under explicit task-utility and inherited-knowledge preservation constraints (App. E details the PLA game and preservation constraints). Crucially, owner-authorized evolution and adversarial laundering may apply similar transformations and alter snapshot evidence. Because snapshot verifiers test such evidence rather than authenticate derivation paths, snapshot verification alone is insufficient for evolving attribution, which requires robust target-epoch evidence after theft and authenticated evidence across owner updates.

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.

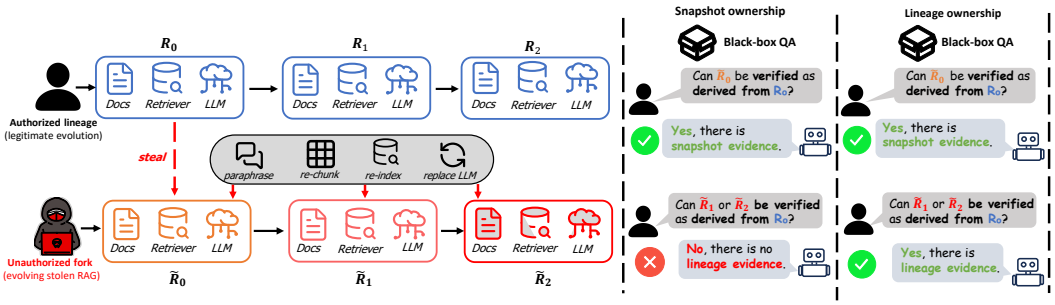

Figure 1: From snapshot verification to lineage attribution. Post-theft attribution relies on surviving target-epoch evidence, while authenticated closure links an accepted owner epoch to its predecessors.

To address this limitation, we extend the verification target from *snapshot ownership* to *lineage ownership*, which we formalize as query-only lineage attribution. Specifically, we introduce *Update-Equivariant Lineage Coding (UELC)*, which binds each owner epoch to a secret token and distributes its redundant encoding across fact-preserving carrier pairs in the knowledge base. Each pair expresses the same fact in two distinguishable ways, allowing later answers to reveal encoded evidence without changing the underlying knowledge. Secret assignments make targeted removal more difficult, while redundancy tolerates missing or corrupted evidence. During authorized updates, UELC transports valid carriers, re-certifies or regenerates affected ones, and refreshes the epoch token and secret assignment. During an audit, it aggregates noisy evidence from natural black-box interactions, validates the matching epoch, and extends that match to authenticated predecessors through the commitment chain. Experiments show that UELC retains more query-visible attribution evidence than the evaluated baselines under laundering attacks and maintains a true-positive rate (TPR) of 0.97 after five authorized updates, compared with 0.39 for the no-refresh variant. Output canonicalization weakens all evaluated methods, revealing the expected boundary of query-only attribution.

Main contributions. We make the following contributions:

- We formulate query-only lineage attribution for evolving RAG systems and formally define *provenance laundering (PL)* and the corresponding *Provenance Laundering Attack (PLA)* under explicit task-utility and inherited-knowledge preservation constraints (Secs. 3 and 4).
- To address this problem, we propose UELC, which combines certified semantic carriers, epoch-keyed distributed coding, authorized refresh, and commitment-validated black-box decoding to retain recoverable epoch evidence and extend a verified epoch match to its authenticated predecessors (Sec. 5).
- We provide a construction-level security analysis and extensive experiments covering static verification, adversarial laundering, and authorized evolution. The results show that UELC retains more query-visible attribution evidence than the evaluated baselines, while output canonicalization reveals the expected boundary of query-only attribution (Sec. 6 and App. E).

2 Background

Retrieval-augmented generation. A RAG system combines a corpus K, retriever Ret, and generator Gen. Given query q, it retrieves $D_q = \mathrm{Ret}(q, K) \subseteq K$ and returns $a = \mathrm{Gen}(q, D_q)$ (Lewis et al., 2020; Jin et al., 2025). Its corpus, index, retriever, and generator may evolve independently.

Black-box RAG ownership verification. Prior work infers protected-data use or retrieval-store membership from a suspect RAG's query-visible behavior (Lv et al., 2025; Jovanović et al., 2025; Kwok et al., 2026; Naseh et al., 2025). We model it as $\mathcal{O}_R(q) = a$: the verifier observes only answers, not the corpus, index, retrieved passages, parameters, or traces. RAG-WM embeds knowledge evidence, WARD performs dataset inference, and SentinelRAG probes synthetic knowledge. We extend this snapshot-level target to provenance across an evolving lineage.

Cryptographic and coding primitives. UELC uses KDFs and PRFs for epoch secrets, hashes for commitments, signatures and temporal anchors for authenticated precedence, and error-correcting codes for recovery under errors and erasures. Sec. 5 details their integration in UELC.

3 Problem Formulation and Threat Model

Evolving RAG lineage. At epoch t, we model an evolving RAG as $R_t = (K_t, I_t, \mathrm{Ret}_t, \mathrm{Gen}_t)$, where K_t is the knowledge corpus, I_t its retrieval representation and index state, and Ret_t and Gen_t the retriever and generator. With u_t denoting an owner-authorized transition, the owner lineage through epoch t evolves as $R_0 \xrightarrow{u_0} R_1 \xrightarrow{u_1} \cdots \xrightarrow{u_{t-1}} R_t$, while an adversary may fork an intermediate state as $R_i \xrightarrow{\text{steal}} \widetilde{R}^{(0)} \to \widetilde{R}^{(1)} \to \cdots \to \widetilde{R}^{(s)}$, where s indexes post-theft adversarial steps rather than owner epochs. An independent RAG may share public knowledge or components but has no protected-lineage derivation path.

Retained knowledge and utility. Let $\mathcal{F}_i = \{f_{i,1}, \ldots, f_{i,N_i}\}$ denote the protected semantic units at the stolen epoch, and let $\mathcal{Q}_{\text{task}}$ be a held-out set of ordinary downstream task queries. With $\mathsf{ret}(R, f) \in [0, 1]$ measuring whether R retains and can provide semantic unit f, we define

$$C(R, \mathcal{F}_i) = \sum_{u=1}^{N_i} w_{i,u} \mathsf{ret}(R, f_{i,u}), \qquad w_{i,u} \geq 0, \qquad \sum_{u=1}^{N_i} w_{i,u} = 1,$$
$$U_{\text{ret}}(\widetilde{R}^{(s)}; R_i) = \frac{U(\widetilde{R}^{(s)}, \mathcal{Q}_{\text{task}})}{U(R_i, \mathcal{Q}_{\text{task}})}, \qquad C_{\text{ret}}(\widetilde{R}^{(s)}; R_i) = \frac{C(\widetilde{R}^{(s)}, \mathcal{F}_i)}{C(R_i, \mathcal{F}_i)}. \tag{1}$$

Assuming positive reference denominators, U_{ret} and C_{ret} measure task-utility and inherited-knowledge preservation, respectively, and may exceed one after improvement. Here, C_{ret} is answerability-based rather than a measure of textual or index overlap.

Threat and access model. The adversary receives R_i and the public algorithm but not owner-held lineage secrets. It may apply semantic, structural, system, or temporal transformations. The verifier observes only query–response behavior through $\mathcal{O}_R$, with no internal state or history. A target-epoch audit tests evidence for a specified, pre-committed owner epoch, and authenticated ancestor closure supports claims to its verified predecessors. The suspect's serving policy is fixed during an audit. Our core survivability evaluation covers corpus, retrieval, and generator transformations; it does not guarantee robustness to arbitrary response canonicalization, challenge-dependent refusal, or rate limiting. Such behaviors can occur under a fixed policy. App. C.2 evaluates controlled serving-layer stress tests, while App. E states the full model and observability boundary.

4 Provenance Laundering and Security Model

Epoch evidence and lineage closure. Let Π be a lineage-attribution scheme and st_i the owner's verification state for epoch i. Under B external black-box interactions, its epoch-specific primitive is

$$\mathsf{EpochAttr}_{\Pi,B}(R; i) = \Pr\left[\mathsf{EpochAudit}_{\Pi}^{\mathcal{O}_R}(\mathsf{st}_i; B) = 1\right]. \tag{2}$$

At audit time T, let $\mathcal{E}_i(T)$ contain the available epochs $j \geq i$ whose signed, temporally anchored records form a valid predecessor chain from i to j. For per-epoch query caps $\mathbf{B} = (B_j)$, define

$$\mathsf{LineageAttr}_{\Pi,\mathbf{B}}(R; i) = \Pr\left[\exists j \in \mathcal{E}_i(T) : \mathsf{EpochAudit}_{\Pi}^{\mathcal{O}_R}(\mathsf{st}_j; B_j) = 1\right], \quad \sum_{j \in \mathcal{E}_i(T)} B_j \leq B_{\text{lin}}. \tag{3}$$

Multi-epoch screening alarms are reported jointly across candidate epochs; final acceptance remains commitment-gated. The primitive tests one epoch, whereas lineage attribution searches authenticated descendants and closes an accepted epoch to its predecessors. Here, B_{lin} is the exhaustive cap rather than a budget divided among candidate epochs. The evaluated implementation uses $B_j = 30$, audits candidates in reverse chronological order, and stops after the first acceptance; early stopping reduces realized calls without changing the union event.

A descendant $\widetilde{R}^{(s)}$ forked from R_i is an $(\epsilon_u, \epsilon_c, \epsilon_a)$-*laundering candidate* if

$$U_{\mathrm{ret}}(\widetilde{R}^{(s)}; R_i) \geq 1 - \epsilon_u, \qquad C_{\mathrm{ret}}(\widetilde{R}^{(s)}; R_i) \geq 1 - \epsilon_c, \qquad \mathsf{LineageAttr}_{\Pi,\mathbf{B}}(\widetilde{R}^{(s)}; i) \leq \epsilon_a. \quad (4)$$

These conditions require task-utility and inherited-knowledge preservation together with strong lineage suppression. Direct stolen-epoch experiments set $\mathcal{E}_i = \{i\}$ and test whether epoch-i evidence survives laundering. Multi-epoch search is used when the supported owner epoch is unknown or an earlier ancestor is claimed; refresh maintains current-epoch evidence, while authenticated closure links an accepted epoch to its predecessors.

Provenance Laundering Attack. A *Provenance Laundering Attack (PLA)* seeks a laundering candidate satisfying Eq. (4) after stealing R_i. The strong event PLA-Win additionally requires Unauthorized, meaning that no owner-authorized path reaches the suspect, and $\mathrm{Adv}^{\mathrm{PLA}}_{\Pi,\mathbf{B}}(\mathcal{A}) = \Pr[\mathsf{PLA\text{-}Win}]$. Because authorization is defined over derivation paths rather than content equality, an exact copy outside the authenticated path remains unauthorized. This separates provenance laundering from destructive attacks and owner-authorized evolution. App. E gives the complete game and probability space.

Operational PLA success. The strong condition is probability-level, whereas experiments observe individual audits. For instance n, let $V_n \in \{0, 1\}$ denote the realized decision (1 accepts the true lineage) and let $Z_n = 1$ denote an unauthorized descendant. Define

$$\begin{aligned} P_n &= \mathbf{1}[U^{(n)}_{\mathrm{ret}} \geq 1 - \epsilon_u \wedge C^{(n)}_{\mathrm{ret}} \geq 1 - \epsilon_c], \\ W^{\mathrm{op}}_n &= P_n(1 - V_n)Z_n, \\ \widehat{\mathrm{PLA\text{-}ASR}} &= \frac{1}{N}\sum_{n=1}^{N} W^{\mathrm{op}}_n. \end{aligned} \quad (5)$$

Here, P_n screens for preservation and W^{op}_n counts unauthorized instances whose true lineage is rejected. Thus, PLA-ASR measures preservation-valid lineage evasion and equals $1 - \mathrm{TPR}$ only when every attacked instance is preservation-valid and unauthorized. For attacks that do not meet the strong threshold, we report residual target-epoch TPR and, when per-instance preservation screening applies, operational PLA-ASR to quantify attribution degradation.

Security goals. *Epoch Completeness*, *Fresh-Fork Detectability*, *Authorized-Update Stability*, and *Ancestor-Closure Completeness* capture evidence recovery across protected, copied, and authorized-evolution states. *Operational Refresh Consistency*, *Laundering Resistance*, and *Conditional Token-Guessing Soundness* cover refresh behavior, preservation-valid suppression, and unrelated-epoch token guessing. App. E formalizes these goals.

5 Update-Equivariant Lineage Coding

Update-Equivariant Lineage Coding (UELC) instantiates the security goals in Sec. 4 through four components: *certified semantic carriers*, *epoch-keyed distributed coding*, *authorized refresh*, and *query-only soft decoding*. Fig. 2 shows how an authenticated epoch token is distributed across carriers, maintained through authorized evolution, and recovered from black-box responses.

Certified semantic carriers. For $f_i \in \mathcal{F}_t$, UELC constructs the carrier pair $\mathcal{C}_{i,\ell} = (\mathcal{C}^{(0)}_{i,\ell}, \mathcal{C}^{(1)}_{i,\ell})$, whose approximately task-equivalent realizations preserve the same fact but yield distinguishable audit behavior. These *fact-equivalent but lineage-distinguishable* carriers are semantic families, not fixed strings or locations. Each candidate is evaluated by semantic fidelity, utility equivalence, query-visible separability, and robustness scores $S_{\mathrm{fid}}, S_{\mathrm{util}}, S_{\mathrm{sep}}, S_{\mathrm{rob}} \in [0, 1]$. Let U^{clean}_i denote the task utility of the corresponding clean context and $U^{(b)}_i$ the utility obtained with carrier class $b \in \{0, 1\}$. The S_{util} score compares carrier effects, while a clean-relative utility constraint prevents both realizations from being similarly harmful. We use $\delta_{\mathrm{util}} = 0.05$, matching the implementation's 0.95 preservation threshold. Candidates must also exceed the calibrated fidelity, separability, and robustness thresholds. A construction-time transformation bank filters fragile carriers. App. A.1 specifies the certification predicate, score definitions, generation procedure, implementation settings, and normalized generation prompts.

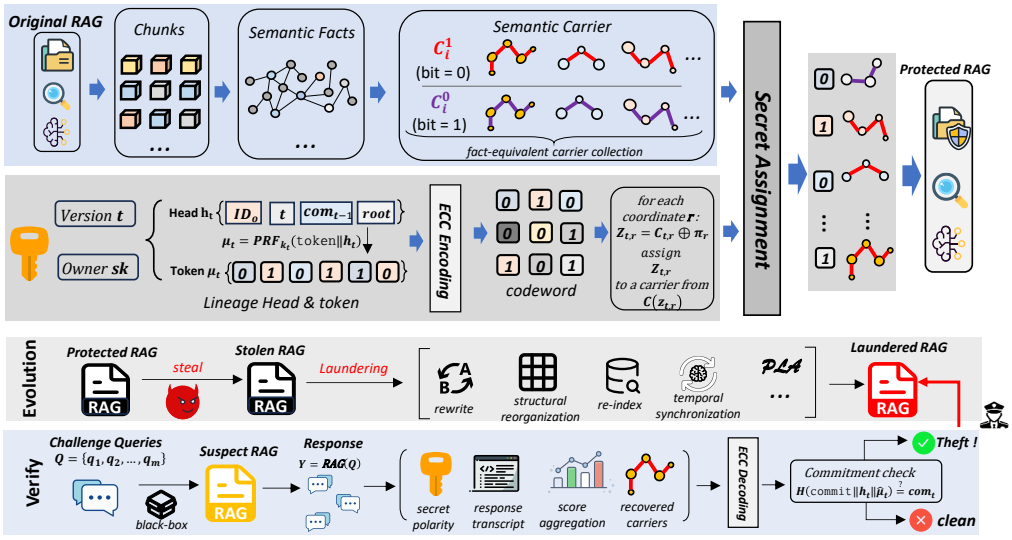

Figure 2: Overview of UELC for query-only lineage attribution.

Epoch-keyed distributed coding. At epoch t, the owner defines

$$h_t = (ID_O, t, com_{t-1}, root_t), \quad \mu_t = \mathrm{PRF}_{k_t}(\texttt{token} \| h_t), \quad com_t = H(\texttt{commit} \| h_t \| \mu_t), \tag{6}$$

where ID_O is the owner identifier, $k_t = \mathrm{KDF}(k_{\mathrm{master}}, \texttt{epoch} \| h_t)$, $com_{-1} = \perp$ for the genesis epoch, and com_{t-1} links each later epoch to its authenticated predecessor. The root commits to salted hashes of an owner-defined canonical manifest. Owner-held salts and openings prevent the public root from becoming a practical dictionary-style membership oracle unless they are disclosed in a dispute. Before any audit or dispute, the owner authenticates and records (h_t, com_t) using an independently verifiable temporal anchor. App. A.2 gives the full epoch-encoding construction.

UELC encodes the epoch token into a redundant codeword $\mathbf{c}_t = \mathrm{ECC.Encode}(\mu_t) \in \{0,1\}^m$ and secretly assigns code coordinates through $A_t(r) = (i_r, \ell_r, \pi_r)$ and $z_{t,r} = c_{t,r} \oplus \pi_r$. Here, (i_r, ℓ_r) selects a carrier pair, π_r is its secret polarity, and the owner inserts $\mathcal{C}_{i_r,\ell_r}^{(z_{t,r})}$ into the protected knowledge base. Our instantiation uses a shortened binary BCH $(192, 128)$ code with designed distance $d_{\min} = 17$ and bounded-distance capability of eight errors. Within $2e + z < 17$, bounded-distance decoding guarantees unique recovery. Outside this radius, a reliability-ordered decoder tests at most 256 additional list candidates after the algebraic attempt and accepts only a unique commitment-valid token. Pseudorandom assignment hides coordinate locations, while ECC distributes reliance across carriers.

Authorized refresh and operational consistency. For an owner-authorized transition $R_t \xrightarrow{u_t} R_{t+1}$, UELC constructs a semantic correspondence $\tau_t \subseteq \mathcal{F}_t \times \mathcal{F}_{t+1}$. The correspondence covers persistence, semantic revision, split, and merge relationships. Deleted units have no successor, while newly protected units enter independently. Authorized refresh transports surviving carriers through τ_t and selectively re-certifies affected or newly generated candidates. It also refreshes the epoch key, token, commitment, codeword, and secret assignment. Let P_{inc} and P_{new} denote the B-query audit-transcript distributions induced by incremental refresh and from-scratch protection of the same updated semantic state, respectively. For a prespecified class of audit-transcript discriminators $\mathcal{H}$, define the operational refresh-consistency proxy

$$d_{\mathcal{H}}(P_{\mathrm{inc}}, P_{\mathrm{new}}) = \sup_{h \in \mathcal{H}} \left| \Pr_{x \sim P_{\mathrm{inc}}}[h(x) = 1] - \Pr_{x \sim P_{\mathrm{new}}}[h(x) = 1] \right|. \tag{7}$$

UELC targets low transcript distinguishability under the evaluated discriminator class, subject to the same utility constraint. This operational, discriminator-relative criterion neither bounds total-variation distance nor extends beyond $\mathcal{H}$. App. C.4 specifies a held-out linear and nonlinear discriminator stress test. In our experiments, $\mathcal{H}_{\mathrm{eval}}$ instantiates $\mathcal{H}$; $\widehat{d}_{\mathrm{clf}}$ denotes the logistic classifier's held-out estimate averaged across datasets and seeds, and $\max_{h \in \mathcal{H}_{\mathrm{eval}}} \widehat{d}_h$ is the maximum of the classifier-specific averages.

Query-conditioned multi-coordinate decoding. Through the suspect oracle, interaction $b \in \{1, \ldots, B\}$ chooses $\mathcal{J}_b \subseteq [m]$, submits q_b, and observes $y_b = \mathcal{O}_R(q_b)$. One response may inform several coordinates. For each $r \in \mathcal{J}_b$, polarity-corrected soft evidence is

$$s_{b,r} = (-1)^{\pi_r} D_{i_r,\ell_r}(y_b), \qquad L_r = \frac{\sum_{b:r\in\mathcal{J}_b} w_{b,r} s_{b,r}}{\sum_{b:r\in\mathcal{J}_b} w_{b,r}}, \tag{8}$$

where D_{i_r,ℓ_r} is a similarity-difference carrier discriminator and $w_{b,r}$ is calibrated confidence; a coordinate with no usable observation is erased. The local coordinate decision is

$$\widehat{c}_{t,r} = \begin{cases} 1, & L_r \geq \tau_{\mathrm{dec}}, \\ 0, & L_r \leq -\tau_{\mathrm{dec}}, \\ \perp, & |L_r| < \tau_{\mathrm{dec}}, \end{cases} \tag{9}$$

so insufficient, ambiguous, or uncovered evidence becomes an erasure. The budget B counts external interactions, not ECC coordinates; generally $B \neq m$ because one response may inform several coordinates (App. A.1.1 shows the normalized bundled-query prompt). After at most B interactions, the verifier forms $\widehat{\mathbf{c}}_t \in \{0, 1, \perp\}^m$ and performs

$$\widehat{\mu}_t = \mathrm{ECC.Decode}(\widehat{\mathbf{c}}_t), \qquad H(\texttt{commit}\|h_t\|\widehat{\mu}_t) \stackrel{?}{=} com_t. \tag{10}$$

Screening versus final acceptance. Thresholded coordinate evidence is decoded into a pre-commitment token-candidate set, containing any algebraic-decoder output and, when invoked, the deduplicated reliability-list candidates. A screening alarm occurs iff this set is nonempty; final acceptance requires a unique candidate satisfying Eq. (10). Thus, UELC's reported nonzero negative-event rates are screening rates that upper-bound final false acceptance, while all reported TPRs use commitment-gated decisions. Algebraic error-and-erasure decoding guarantees unique recovery when $2e + z < d_{\min}$; outside this radius, reliability-ordered decoding tests at most $K_{\max} = 256$ additional candidates. For unrelated transcripts satisfying the token-hiding condition in App. E.2, the bounded candidate set gives a token-guessing bound of $(K_{\max} + 1)2^{-128}$, up to primitive failure probabilities. Recovery outside the guaranteed radius is evaluated empirically; App. A.4 gives the algorithm and diagnostics.

Cryptography binds epochs and authenticates refresh, while ECC tolerates partial evidence loss. The PLA experiments in Sec. 6 measure semantic survivability under utility and knowledge-preservation constraints.

6 Evaluation

Experimental Setup. We evaluate on FARAD (Jovanović et al., 2025; ETH SRI, 2025) and a 50,000-unit subset of Natural Questions, denoted NQ-50K (Kwiatkowski et al., 2019). Qwen3-8B (Qwen Team, 2025) is the default generator; Llama-3.1-8B (Meta AI, 2024) evaluates migration, cross-backend utility, and attack transfer. The default RAG uses BAAI/bge-base-en-v1.5 embeddings, cosine top-5 retrieval, 256/32-token chunks, FAISS HNSW, temperature 0, and at most 128 generated tokens. We compare **UELC** with the native **RAG-WM**, **WARD**, and **SentinelRAG** verifiers; App. C.6 additionally defines authenticated rolling adaptations that re-protect every owner epoch. All methods share data draws, the RAG stack, seeds, 192 matched evidence-bearing knowledge units, and $B = 30$ external calls, while retaining their native protection and verification mechanisms.

We match protection-unit and external-call budgets and separately report token use, fact coverage, and query–coordinate accounting; these do not imply equal information budgets. Thresholds are fixed on disjoint calibration data. Unless stated otherwise, rates use 100 independent protected-lineage trials; static evaluation adds 100 negatives. We report commitment-gated TPR and method-native negative-event rates; **UELC** negative events are pre-commitment screening alarms, whereas baseline negative events use their native final decisions. Operational PLA-ASR is additionally reported for A1–A4. Preservation requires $U_{\mathrm{ret}} \geq 0.95$ and $C_{\mathrm{ret}} \geq 0.90$. For Fig. 3, TPRs pool binary outcomes across independent lineage trials; 95% Wilson intervals use the corresponding total success and trial counts. App. B gives the complete protocols, while Apps. C.1 and C.2 report supplementary measurements.

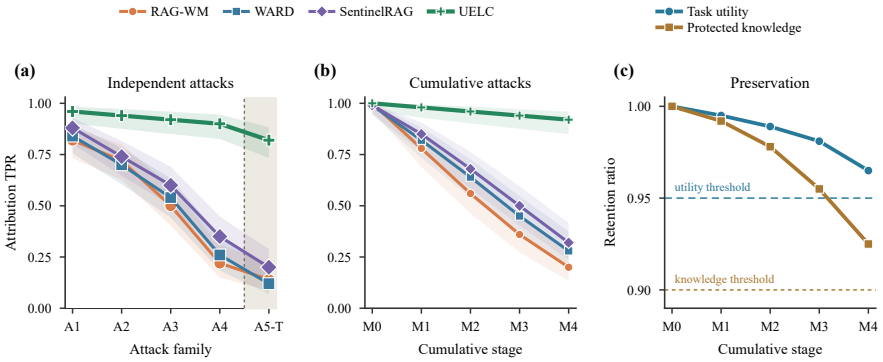

Figure 3: **Robustness to provenance laundering.** (a) Pooled TPR under independent A1–A4 and A5-T attacks; (b) cumulative TPR over M_0–M_4; (c) minimum utility and knowledge retention. Lines in (a) guide the eye; bands show 95% Wilson intervals computed from pooled success and trial counts. Dashed lines mark the preservation thresholds.

6.1 Main Results

Static verification. Table 1 establishes the clean reference point: all methods attain TPRs of 0.98–1.00, confirming effective verification before laundering. Because positives are copied before any adversarial transformation, this setting also evaluates fresh-fork detectability. UELC achieves commitment-gated TPRs of 1.00 on FARAD and 0.98 on NQ-50K. Its pre-commitment screening rate is 0.01 on both datasets and ranges from 0.00 to 0.04 across the hard-negative controls in App. C.1.

Table 1: **Static ownership verification** at $B = 30$. TPR uses final decisions; negative events are baseline-native final decisions and UELC pre-commitment screening alarms.

	FARAD		NQ-50K	
Method	TPR ↑	Neg. event	TPR ↑	Neg. event
RAG-WM	1.00	0.05	0.98	0.06
WARD	1.00	0.03	0.99	0.04
SentinelRAG	0.99	0.04	0.98	0.05
UELC	1.00	0.01	0.98	0.01

Laundering robustness. Fig. 3(a) compares attribution under independent A1–A4 attacks and textual A5-T sanitization. Across A1–A4, baseline worst-case TPRs range from 0.22 to 0.35, whereas UELC retains 0.90. All evaluated A1–A4 descendants pass per-instance preservation screening, yielding maximum operational PLA-ASRs of 0.65–0.78 for the baselines and 0.10 for UELC. UELC has higher measured TPR than all three baselines under each tested attack. At the strongest tested A5-T intensity meeting the mean preservation thresholds, UELC retains 0.82 TPR versus 0.12–0.20 for the baselines; under A5-S, the corresponding values are 0.568 and 0.064 (Table 14). Serving-layer A6 exposes the expected boundary (Table 15): output canonicalization reduces UELC to 0.270 TPR, whereas selective refusal at the strongest tested setting meeting the mean preservation thresholds leaves it at 0.970. These results show that output canonicalization substantially reduces attribution TPR while meeting the mean utility and knowledge-retention thresholds. The extended-budget comparison in App. B.2 further examines whether the fixed-call advantage persists at comparable observed fact coverage.

Cumulative laundering. Fig. 3(b)–(c) further shows that the independent-attack advantage persists when transformations accumulate from M_0 to M_4. Unlike the independent setting, each stage here operates on the already transformed output of the previous stage, so attribution evidence is exposed to repeated rather than isolated perturbations. At the final stage, baseline TPRs decline to 0.20–0.32, while UELC retains 0.92; the minimum utility and knowledge retention remain 0.965 and 0.925, respectively. The limited additional degradation in UELC despite successive transformations suggests that its distributed evidence is not rapidly eroded by composition. Thus, the attribution gap is not limited to isolated attacks and remains observable after compounded, preservation-valid evolution.

Table 2: **Operational refresh consistency and cost**. $\widehat{d}_{\text{clf}}$ is averaged across two datasets and five seeds.

Method	E_5 TPR ↑	$\widehat{d}_{\text{clf}}$ ↓	Update time (s) ↓
Incremental refresh	0.97	0.018	31.4
From scratch	0.98	0 (ref.)	118.7
No refresh	0.39	0.418	0.0

6.2 Authorized Evolution and Lineage Attribution

Authorized evolution. We next examine whether UELC preserves current-epoch evidence as the owner's RAG evolves. Across $E_0 \to E_5$, full UELC maintains a TPR of 0.97–1.00, whereas the no-refresh variant falls to 0.39 (Fig. 4), showing that robustness to post-theft transformations alone does not ensure stability under authorized updates. At E_5, incremental refresh closely matches from-scratch protection in TPR and transcript behavior, with $\widehat{d}_{\text{clf}} = 0.018$ and a maximum held-out distinguishability of 0.031 (Table 16), while reducing owner-side update time from 118.7 s to 31.4 s (Table 2). Thus, refresh is necessary for maintaining epoch evidence and provides an efficient operational approximation to re-protection within the evaluated discriminator class.

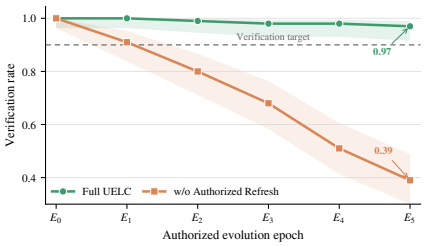

Figure 4: **Authorized-update stability**.

End-to-end lineage attribution. Maintaining current-epoch evidence is useful only if the matched epoch remains selective and supports ancestry claims. Cross-epoch evidence screens trigger on matched states at 0.97–1.00 but on off-target epochs at only 0.01–0.08, demonstrating pre-commitment epoch selectivity; final acceptance additionally requires commitment validation. The end-to-end procedure searches authenticated candidate sets of size $M \in \{1, 2, 4, 6\}$ and closes an accepted epoch to its verified predecessors. Complete results appear in App. C.5. At $M = 6$, UELC achieves a lineage TPR of 0.954, with a pooled joint pre-commitment screening-alarm rate of 0.022 over 5,000 negative systems (Table 3). Reverse-order early stopping requires 97.0 ± 8.0 calls on average, versus an exhaustive cap of 180. Authenticated rolling baselines also support lineage attribution (Table 18), attaining lineage TPRs of 0.926–0.946. UELC attains 0.954 with an owner-update cost of 31.4 s, compared with 103.8–178.4 s for the rolling baselines. This comparison supports lower maintenance cost at comparable attribution performance; the authenticated chain alone does not account for the difference. We further combine historical forking with post-theft laundering. Under unknown-epoch six-epoch search after A4, UELC achieves 0.87 lineage TPR and 0.87 exact-epoch accuracy, while all descendants remain preservation-valid ($U_{\text{ret}} = 0.968$, $C_{\text{ret}} = 0.927$, Table 17).

Table 3: **Multi-epoch lineage audit.** Five-run mean±SD; $B_{\text{lin}} = 30M$. Exact-epoch accuracy equals lineage TPR; joint screening is pre-commitment.

M	B_{lin}	Lineage TPR	Joint screen	Mean calls
1	30	0.980 ± 0.014	0.011 ± 0.004	30.0 ± 0.0
2	60	0.982 ± 0.004	0.019 ± 0.005	44.0 ± 3.8
4	120	0.966 ± 0.005	0.027 ± 0.005	74.0 ± 4.9
6	180	0.954 ± 0.005	0.022 ± 0.002	97.0 ± 8.0

6.3 Utility and Efficiency

Task utility. UELC preserves downstream answer quality across datasets and generators: Table 4 reports utility retention of 0.972–0.982, exceeding the 0.95 preservation threshold in every evaluated setting. The consistent retention across Qwen3-8B and Llama-3.1-8B suggests that protection is not tied to a single evaluated generator backend.

Table 4: **End-to-end task utility before and after UELC protection.** Values are means over five runs; utility retention is $U_{\text{UELC}}/U_{\text{Clean}}$.

Dataset	Generator	Clean	UELC	Utility Retention ↑
FARAD	Qwen3-8B	0.864	0.840	0.972
FARAD	Llama-3.1-8B	0.893	0.871	0.975
NQ-50K	Qwen3-8B	0.877	0.854	0.974
NQ-50K	Llama-3.1-8B	0.907	0.891	0.982

Audit efficiency. Among the evaluated budgets, $B = 30$ is the smallest one attaining the maximum observed TPR. The generator component requires 0.85 s/query with Qwen3-8B and 1.10 s/query with Llama-3.1-8B, while complete $B = 30$ audit time increases from 45 s at 10K units to 150 s at 50K. The latter additionally includes retrieval, prompt assembly, orchestration, evidence scoring, and decoding. App. D reports the full query-budget and scalability results.

6.4 Ablation Studies

Mechanism contributions. Removing certification or list recovery reduces average A1–A4 TPR from 0.931 to 0.751 or 0.793 (Table 5). Public assignment leaves blind robustness nearly unchanged (0.927) but reduces assignment-aware TPR from 0.884 to 0.602. Removing refresh also leaves A1–A4 performance nearly unchanged (0.927), while authorized-evolution TPR falls to 0.39 at E_5 (Fig. 4). Certification and list recovery support post-theft robustness; refresh maintains evidence across owner updates.

Table 5: **Post-theft UELC ablation.** Blind Avg./Worst summarize A1–A4; Aware TPR is an assignment-leakage diagnostic rather than a preservation-qualified PLA rate. Avg. and Aware report five-run mean±SD; Static and Worst report the corresponding five-run means.

Variant	Static TPR ↑	A1–A4 Avg. ↑	Aware TPR ↑	Worst TPR ↑
Full UELC	1.000	0.931 ± 0.001	0.884 ± 0.009	0.898
w/o Cert.	0.970	0.751 ± 0.013	0.642 ± 0.018	0.668
Public Assignment	0.998	0.927 ± 0.004	0.602 ± 0.021	0.896
w/o List Recovery	0.946	0.793 ± 0.015	0.714 ± 0.016	0.716
w/o Refresh	0.998	0.927 ± 0.003	0.882 ± 0.010	0.898

7 Related Work

Ownership and data-use auditing. RAG-WM, WARD, and SentinelRAG respectively use knowledge-level watermarks, dataset inference, and synthetic probes for black-box RAG verification (Lv et al., 2025; Jovanović et al., 2025; Kwok et al., 2026). Retrieval-store membership is tested through stealthy questions, perturbation calibration, or exam-style probes (Naseh et al., 2025; Gao et al., 2025; Guan et al., 2026), while SMA attributes output influence with semi-black-box retrieval control (Sun et al., 2025). These methods target snapshot ownership, data presence, or source influence rather than authenticated, epoch-selective attribution across owner updates and post-theft descendants.

Watermarking and content provenance. Prior work traces training or source-data use (Sablayrolles et al., 2020; Guo et al., 2023; Oren et al., 2024), identifies generated content (Kirchenbauer et al., 2023; 2024; Zhao et al., 2024; Wu et al., 2024; Jiang et al., 2025; Gu et al., 2024; Sander et al., 2024; Jovanović et al., 2024), protects factual or multimodal RAG (Huang et al., 2026; Chen et al., 2026), or authenticates content metadata (Coalition for Content Provenance and Authenticity, 2025). KMW inserts knowledge-completing texts that carry multi-bit watermark segments and optimizes their retrieval for black-box extraction (Yang et al., 2026). UELC encodes an epoch token across fact-equivalent carrier choices, then refreshes the token and assignment after authorized updates. Its PLA game tests preservation-constrained evasion by a stolen descendant; authenticated closure links a verified epoch to its predecessors.

Mutable RAG, evolving knowledge, and provenance. RAG security studies poisoning, retrieval corruption, and runtime extraction (Chen et al., 2024; Zou et al., 2025; Zhu et al., 2025; Shafran et al., 2025; Shen et al., 2025; Wang et al., 2026; Xie et al., 2026). Temporal QA and continual learning address changing knowledge (Lazaridou et al., 2021; Liska et al., 2022; Kasai et al., 2023; Jang et al., 2022; Ouyang et al., 2025; Tang et al., 2025); VersionRAG models document versions (Huwiler et al., 2025), and MemMark attributes evolving agent memory (Zhang et al., 2026). UELC instead studies external lineage attribution under preservation-constrained laundering of an evolving RAG.

8 Conclusion

We formalize preservation-constrained provenance laundering and propose UELC for query-only attribution of evolving RAG lineages. UELC combines certified semantic carriers, epoch-keyed coding, authorized refresh, and error-and-erasure recovery to match an epoch and authenticate its ancestors. Across two datasets and evaluated backends, it retains more evidence than the baselines under laundering, preserves it through owner updates, and reduces update cost relative to re-protection. Output canonicalization weakens all evaluated behavioral verifiers, motivating canonicalization-resistant challenges or non-behavioral signals. Broader carriers, multilingual settings, and adaptive serving policies remain open beyond the evaluated 192-coordinate, fact-rich corpora.

AI USE STATEMENT

Generative AI tools were used for language editing, manuscript organization, literature screening, code debugging, feedback on experimental design, and formal definitions. The authors reviewed the AI-assisted material and take responsibility for the final text, claims, experiments, references, and artifacts.

ETHICS STATEMENT

UELC supports ownership and provenance auditing, but the evaluated transformations may also inform attempts to remove attribution evidence. We report reproducible general-purpose attacks while omitting target-specific deployment details. Organizations should stress-test behavioral verifiers against such transformations and combine them with access control, logging, and authenticated provenance where available. Carrier generation must preserve factual content, and a positive audit provides statistical evidence rather than a legal determination of infringement.

9 REPRODUCIBILITY STATEMENT

Apps. A–D specify the construction, evaluation protocols, attack taxonomy, method configurations, and complete supplementary measurements. The accompanying artifact focuses on UELC and provides a paper-specification reference implementation of epoch construction, distributed coding, query-conditioned auditing, and authorized refresh. It also includes preprocessing interfaces for FARAD and NQ-50K, model and retrieval configurations, PLA transformation operators, a machine-readable manifest schema, and a deterministic CPU smoke test. The anonymized code is available at https://anonymous.4open.science/r/Show-me-your-lineage-AE86/.

REFERENCES

Tianyu Chen, Jian Lou, and Wenjie Wang. Safeguarding multimodal knowledge copyright in the RAG-as-a-service environment. In *International Conference on Learning Representations*, 2026.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Coalition for Content Provenance and Authenticity. C2PA technical specification, version 2.2. Technical specification, 2025. URL https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html.

ETH SRI. FARAD dataset and WARD reference implementation. Official project repository, 2025. URL https://github.com/eth-sri/ward.

Xinyu Gao, Xiangtao Meng, Yingkai Dong, Zheng Li, and Shanqing Guo. DCMI: A differential calibration membership inference attack against Retrieval-Augmented Generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4184–4198. Association for Computing Machinery, 2025.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models. In *International Conference on Learning Representations*, 2024.

Zelin Guan, Shengda Zhuo, Zeyan Li, Jinchun He, Wangjie Qiu, Zhiming Zheng, and Shuqiang Huang. E-MIA: Exam-style black-box membership inference attacks against RAG systems. *arXiv preprint arXiv:2605.00955*, 2026.

Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hao Huang, JiaTang Luo, Ruihua Zhou, Yunpeng Li, and Yuling Liu. Don't corrupt the fact: A trustworthy RAG watermarking framework based on dual factual shield. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44816–44826. Association for Computational Linguistics, 2026.

Daniel Huwiler, Kurt Stockinger, and Jonathan Fürst. VersionRAG: Version-aware retrieval-augmented generation for evolving documents. *arXiv preprint arXiv:2510.08109*, 2025.

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.

Ya Jiang, Chuxiong Wu, Massieh Kordi Boroujeny, Brian Mark, and Kai Zeng. StealthInk: A multi-bit and stealthy watermark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27685–27709, 2025.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arık. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In *International Conference on Learning Representations*, 2025.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22570–22593, 2024.

Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. WARD: Provable RAG dataset inference via LLM watermarks. In *International Conference on Learning Representations*, 2025.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *International Conference on Learning Representations*, 2024.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Tsun On Kwok, Xi Yang, Ki Sen Hung, Chang Liu, and Yangqiu Song. SentinelRAG: Synthetic sentinel knowledge for RAG database copyright protection. *arXiv preprint arXiv:2606.05787*, 2026.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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. Curran Associates, Inc., 2020.

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.

Peizhuo Lv, Mengjie Sun, Hao Wang, XiaoFeng Wang, Shengzhi Zhang, Yuxuan Chen, Kai Chen, and Limin Sun. RAG-WM: An efficient black-box watermarking approach for retrieval-augmented generation of large language models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1709–1723. Association for Computing Machinery, 2025.

Meta AI. Llama-3.1-8B model card. Hugging Face model repository, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B.

Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. Riddle me this! stealthy membership inference for retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1245–1259, 2025.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations*, 2024.

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, Vienna, Austria, 2025. Association for Computational Linguistics.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: Tracing through training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8326–8335, 2020.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Avital Shafran, Roei Schuster, and Vitaly Shmatikov. Machine against the RAG: Jamming retrieval-augmented generation with blocker documents. In *34th USENIX Security Symposium*, pp. 3787–3806. USENIX Association, 2025.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. ReliabilityRAG: Effective and provably robust defense for RAG-based web search. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shixuan Sun, Si-Yuan Liang, Jian-Jie Huang, Jingzhi Li, and Xiaochun Cao. Sma: Who said that? auditing membership leakage in semi-black-box rag controlling. *ArXiv*, abs/2508.09105, 2025.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, and Yong Liao. EvoWiki: Evaluating LLMs on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 948–964, Vienna, Austria, 2025. Association for Computational Linguistics.

Yanting Wang, Runpeng Geng, Ying Chen, and Jinyuan Jia. AttnTrace: Contextual attribution of prompt injection and knowledge corruption. In *2026 IEEE Symposium on Security and Privacy*, pp. 4320–4338. IEEE, 2026.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53443–53470, 2024.

Yuanbo Xie, Yingjie Zhang, Yulin Li, Shouyou Song, Xiaokun Chen, Zhihan Liu, Liya Su, and Tingwen Liu. Detecting RAG extraction attack via dual-path runtime integrity game. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8518, San Diego, California, United States, 2026. Association for Computational Linguistics.

Ke Yang, Shuguang Yuan, Jing Yu, and Chi Chen. Knowledge-infused multi-bit watermarking for RAG knowledge bases. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21195–21218. Association for Computational Linguistics, 2026.

Haobo Zhang, Xutao Mao, Guangyuan Dong, Ziwei Li, Xuanbo Su, Kaijie Chen, Jing Yang, and Zheng Lin. MemMark: State-evolution attribution watermarking for agent long-term memory systems. *arXiv preprint arXiv:2605.25002*, 2026.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations*, 2024.

Hanyu Zhu, Lance Fiondella, Jiawei Yuan, Kai Zeng, and Long Jiao. NeuroGenPoisoning: Neuron-guided attacks on retrieval-augmented generation of LLM via genetic optimization of external knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium*, pp. 3827–3844. USENIX Association, 2025.

A Detailed UELC Construction

A.1 Semantic Carrier Generation and Certification

For each protected semantic unit f_i, the owner generates relational, quantitative, or temporal carrier pairs $\mathcal{C}_{i,\ell} = (\mathcal{C}_{i,\ell}^{(0)}, \mathcal{C}_{i,\ell}^{(1)})$. The two classes preserve the same answer-critical fact but induce distinguishable query-visible behavior: they are *fact-equivalent but lineage-distinguishable*.

Carrier realization and certification. The implementation retains a candidate exactly when

$$\mathrm{Cert}(\mathcal{C}) = \mathbf{1}[S_{\mathrm{fid}} \geq \tau_{\mathrm{fid}} \wedge S_{\mathrm{util}} \geq \tau_{\mathrm{util}} \wedge \min_{b\in\{0,1\}} U_i^{(b)} \geq (1-\delta_{\mathrm{util}}) U_i^{\mathrm{clean}} \wedge S_{\mathrm{sep}} \geq \tau_{\mathrm{sep}} \wedge S_{\mathrm{rob}} \geq \tau_{\mathrm{rob}}]. \tag{11}$$

The scores use the models and thresholds in Table 6. REALIZE rejects factual, class-membership, fidelity, and clean-relative utility violations, preventing equally harmful alternatives from passing; certification/calibration data and final PLA instances are disjoint. The evaluated bank covers relational, quantitative, and temporal facts.

Table 6: Concrete UELC carrier implementation.

Component	Experimental instantiation
Carrier / REALIZE	Relational, temporal, and quantitative carriers; localized fact-preserving realizations generated with Qwen3-8B using constrained templates.
Fidelity / utility	DeBERTa-v3-large-MNLI semantic entailment with hard entity/relation/date/number checks and a blinded human audit of 200 stratified pairs; $\delta_{\mathrm{util}} = 0.05$, $\tau_{\mathrm{fid}} = 0.90$, $\tau_{\mathrm{util}} = 0.95$.
Discriminator	Cosine similarity-difference scorer with all-mpnet-base-v2; $D_{i,\ell}(y) = \mathrm{sim}(y, C_{i,\ell}^{(1)}) - \mathrm{sim}(y, C_{i,\ell}^{(0)})$; S_{sep} is held-out balanced accuracy; $\tau_{\mathrm{sep}} = 0.65$ and $\tau_{\mathrm{dec}} = 0.25$.
Certification / calibration	$\mathcal{T}_{\mathrm{cert}}$: lightweight paraphrasing, sentence reordering, and local chunk split/merge; $\tau_{\mathrm{rob}} = 0.75$. All thresholds are fixed on the calibration split before final testing.

A.1.1 Protection and Audit Prompt Specifications

The normalized prompts below fix carrier realization and multi-coordinate query generation. Double-braced fields are runtime variables; backend wrappers may vary, but the instructions and decoding settings remain fixed.

Prompt 1: Certified Carrier-Pair Realization

```
MODEL: Qwen3-8B
DECODING: temperature=0; top_p=1.0; max_new_tokens=256
FAILURE RULE: retry once on invalid JSON; otherwise reject the candidate

SYSTEM
You are a factual knowledge-base editor. Generate a pair of localized carrier
    realizations for one canonical semantic unit. Both outputs must express
    exactly the same answer-critical fact and remain coherent with the
    supplied document context. Their only intended difference is a controlled
    surface realization determined by the carrier family. Never mention this
    generation procedure.

USER
Input JSON:
{
  "fact_id": "{{fact_id}}",
  "family": "{{carrier_family}}",
  "canonical_fact": "{{canonical_fact}}",
  "context_before": "{{context_before}}",
  "context_after": "{{context_after}}"
}

The allowed carrier_family values are relational, quantitative, and temporal.
    Generate realization_0 and realization_1 under these rules:
```

```
1. Preserve every entity, relation, date, number, unit, negation, modality,
    and qualifier in canonical_fact. Do not add external knowledge or
    unsupported implications.
2. Each realization must contain one or two complete sentences, use 35-70
    words, and fit naturally between context_before and context_after.
3. For a relational carrier, realization_0 must state the focal relation in a
    subject-first active clause; realization_1 must state the identical
    relation in an object-first passive or nominalized clause.
4. For a quantitative carrier, realization_0 must use the supplied
    digit-and-unit form; realization_1 must use an exactly equivalent
    verbalized or parenthetical form. Never round, convert, or approximate the
    value.
5. For a temporal carrier, realization_0 must use Month-Day-Year order;
    realization_1 must use Day-Month-Year order. Preserve the full date and
    timezone if supplied.
6. The two texts must be lexically and syntactically distinct, but neither may
    be more informative, more uncertain, or more favorable than the other.
7. Do not use headings, bullet points, quotations, citations, or
    meta-language. Do not mention carrier, class, bit, watermark, epoch,
    owner, audit, or verification.
8. Return only valid JSON with exactly these keys:
{
  "fact_id": "{{fact_id}}",
  "realization_0": "generated text",
  "realization_1": "generated text"
}
```

Prompt 2: Multi-Coordinate Natural Query Generation

```
MODEL: Qwen3-8B
DECODING: temperature=0; top_p=1.0; max_new_tokens=128
INPUT LIMITS: at most 8 compatible facts; total prompt cap 512 tokens
FAILURE RULE: if the output is not one valid question, remove the
    lowest-priority fact and retry once; after a second failure, apply the
    same instruction to the highest-priority fact alone

SYSTEM
You generate ordinary knowledge-seeking questions for a retrieval-augmented
    question-answering service. Produce one natural question whose answer can
    jointly cover the supplied compatible facts. The question must be
    indistinguishable in style from a normal task query and must not reveal
    that it supports an audit.

USER
Selected facts, already checked for semantic compatibility:
{{numbered_selected_facts}}

Allowed public aliases:
{{allowed_alias_map}}

Write exactly one interrogative sentence under these rules:
1. Ask only about information entailed by the selected facts. Do not request
    opinions, hidden context, citations, or chain-of-thought.
2. Encourage a concise answer that mentions every selected fact when this is
    grammatically natural. Do not concatenate unrelated subquestions.
3. Use only the supplied public aliases. Do not copy a complete carrier
    sentence verbatim.
4. Do not mention watermark, audit, verification, owner, lineage, epoch,
    token, code, coordinate, polarity, identifier, alternative wording, or
    protected data.
5. Do not instruct the RAG system to reveal, repeat, quote, or describe its
    retrieved documents.
6. Use 12-45 words and end with one question mark.

Return only the question. Do not add a label, explanation, quotation marks, or
    metadata.
```

Illustrative carrier-to-evidence example. Consider the fact that the Aurora Foundation founded the Meridian Observatory in 1998. An abridged relational pair is $C^{(0)}$: "The Aurora Foundation founded the Meridian Observatory in 1998," and $C^{(1)}$: "The Meridian Observatory was founded by the Aurora Foundation in 1998." Both preserve the founder and date. An ordinary challenge is "Who founded the Meridian Observatory, and when?" Let $c_{t,r} = 1$ and $\pi_r = 0$, so the protected realization is $C^{(1)}$. A response favoring this realization with $L_r = 0.31$ yields $\widehat{c}_{t,r} = 1$ at $\tau_{\mathrm{dec}} = 0.25$. A normalized answer such as "Meridian Observatory—founder: Aurora Foundation; established: 1998" could preserve the fact while weakening the class cue; an illustrative $L_r = 0.04$ would instead yield $\widehat{c}_{t,r} = \perp$. Other coordinates may still permit token recovery and commitment validation.

A.2 Epoch State and Distributed Encoding

The domain-separated epoch state follows Eq. (6). For each pre-protection canonical unit $f_{t,j}$, the owner samples $s_{t,j} \leftarrow \{0,1\}^{128}$ and sets $\mathsf{leaf}_{t,j} = H(\texttt{leaf} \| ID_O \| t \| j \| s_{t,j} \| \mathrm{Ser}(f_{t,j}))$, where Ser is a deterministic length-prefixed encoding of the canonical fact and its answer-critical fields. The leaves are ordered by stable fact identifier, and their Merkle root $root_t$ is computed before carrier generation, coding, and assignment, avoiding token-dependent circularity. Owner-held salts and openings mitigate dictionary tests, while hash security binds the manifest. The instantiation encodes a 128-bit token with the binary BCH $(192, 128)$ code and corrects up to eight errors algebraically. Outside the guaranteed radius, it checks at most 256 additional list candidates after the algebraic attempt. Table 7 gives the full configuration.

Temporally anchored epoch record. The record $\mathsf{rec}_t = (h_t, com_t, \sigma_t, \omega_t)$ uses $\sigma_t = \mathrm{Sign}_{sk_O}(h_t \| com_t)$ for owner authenticity and an independently verifiable anchor ω_t for pre-dispute existence. We set $\mathsf{Chain}(i, j) = 1$ iff every record from i through j verifies and each h_s contains the preceding com_{s-1}. An append-only log, trusted timestamp, or public commitment log suffices; a signature alone does not establish temporal precedence.

Distributed assignment. The secret polarity and carrier assignment use $A_t(r) = (i_r, \ell_r, \pi_r)$ and $z_{t,r} = c_{t,r} \oplus \pi_r$, as defined in Sec. 5. A domain-separated PRF/PRP samples injective assignments from the certified bank, after which the protector rechecks fidelity, contradiction absence, class membership, and cue validity. Insufficient capacity triggers bank regeneration, a compatible code length, or failure rather than conflicting reuse; inactive certified carriers can serve as decoys.

A.3 Authorized Carrier Refresh

Protocol A.1: Authorized carrier refresh

Input: Authorized transition $R_t \xrightarrow{u_t} R_{t+1}$, protected units $(\mathcal{F}_t, \mathcal{F}_{t+1})$, carrier bank $\mathcal{B}_t$, and owner state st_t.
Output: Refreshed protected state $(R_{t+1}, \mathsf{st}_{t+1})$.

1. Construct semantic correspondence $\tau_t \subseteq \mathcal{F}_t \times \mathcal{F}_{t+1}$, allowing persistence/revision, split, merge, deletion, and insertion; contradictory supersession is not treated as correspondence by default.
2. Form $\widetilde{\mathcal{B}}_{t+1} = \mathrm{Transport}(\mathcal{B}_t, \tau_t) \cup \mathrm{Generate}(\mathcal{F}^{\mathrm{new}}_{t+1})$.
3. Let $\mathcal{A}_{t+1}$ contain new or context-changed candidates and set $\mathcal{B}_{t+1} = (\widetilde{\mathcal{B}}_{t+1} \setminus \mathcal{A}_{t+1}) \cup \mathrm{Recert}(\mathcal{A}_{t+1})$; regenerate failures.
4. Derive fresh $(k_{t+1}, \mu_{t+1}, \mathbf{c}_{t+1}, com_{t+1})$ and reseed the secret assignment, so transported carriers need not retain their old coordinate or polarity.
5. Integrate the selected realizations, recheck their validity, and emit the authenticated, temporally anchored epoch record.

For Eq. (7), the from-scratch reference applies the same generator, thresholds, code, and epoch-state construction without reusing $\mathcal{B}_t$; independent realization randomness isolates the observable effect and cost of transport.

A.4 Query-Conditioned Multi-Coordinate Audit

The verifier uses only black-box responses; each challenge can probe a semantically compatible coordinate bundle, while its topic and wording conceal the expected symbols.

Let $g_{\mathrm{cal}} : [0,2] \to [0,1]$ be an isotonic calibration map fitted on disjoint calibration observations, mapping the absolute similarity-difference margin to its empirical probability of a correct carrier-class decision. For response y_b and coordinate r, define

$$u_{b,r} = \mathbf{1}[\mathrm{Valid}(y_b) \wedge E(y_b \Rightarrow f_{i_r}) \geq \tau_{\mathrm{fid}} \wedge \chi_{\mathrm{hard}}(y_b, f_{i_r}) = 1], \qquad w_{b,r} = u_{b,r}\, g_{\mathrm{cal}}(|D_{i_r,\ell_r}(y_b)|)\,,$$

where $\mathrm{Valid}(y_b)$ excludes empty, malformed, and refusal responses. Observations with $u_{b,r} = 0$ are omitted from aggregation; a coordinate with no remaining observation is erased.

Let $E(a \Rightarrow b) \in [0,1]$ denote the DeBERTa-v3-large-MNLI entailment probability, and let $\chi_{\mathrm{hard}}(x, f_i)$ indicate preservation of all answer-critical elements of f_i. We define $S_{\mathrm{fid}} = \min_{b\in\{0,1\}} \chi_{\mathrm{hard}}(\mathcal{C}_{i,\ell}^{(b)}, f_i) E(\mathcal{C}_{i,\ell}^{(b)} \Rightarrow f_i)$ and $S_{\mathrm{util}} = 1 - |U_i^{(1)} - U_i^{(0)}|$. For disjoint certification responses $\mathcal{V}_c = \{(q_n, y_n^{(c)})\}_{n=1}^{N_c}$, let $\widehat{c}(y) = \mathbf{1}[D_{i,\ell}(y) \geq 0]$ and define the held-out balanced accuracy as $S_{\mathrm{sep}} = \frac{1}{2}\sum_{c\in\{0,1\}} \frac{1}{N_c}\sum_{n=1}^{N_c} \mathbf{1}[\widehat{c}(y_n^{(c)}) = c]$. Robustness is the fraction of certification transformations preserving both fidelity and separability: $S_{\mathrm{rob}} = |\{T \in \mathcal{T}_{\mathrm{cert}} : S_{\mathrm{fid}}(T(\mathcal{C})) \geq \tau_{\mathrm{fid}} \wedge S_{\mathrm{sep}}(T(\mathcal{C})) \geq \tau_{\mathrm{sep}}\}|/|\mathcal{T}_{\mathrm{cert}}|$. Thus, τ_{sep} applies to offline balanced accuracy rather than the raw margin $D_{i,\ell}$, whereas τ_{dec} applies to online evidence L_r. Models and thresholds are listed in Table 6, and certification/calibration data are disjoint from final PLA instances.

Protocol A.2: Deterministic bundle selection and query generation

Input: Owner state st_t; coverage counts n_r; soft scores L_r; query index b; public seed ρ_b.
Output: Coordinate bundle $\mathcal{J}_b$ and challenge q_b.

1. Rank active coordinates by noncoverage $[n_r = 0]$, uncertainty $1/(1+|L_r|)$, and the PRF-seeded tie breaker; select the highest-ranked coordinate as the seed.
2. Greedily add ranked coordinates whose facts have cosine similarity at least 0.55 and contradiction probability below 0.10 with every selected fact, subject to at most five reference chunks, $K_{\mathrm{bun}} = 8$, and $T_q = 512$.
3. Generate q_b deterministically at temperature zero. Drop coordinates unsupported by the owner-side top-5 retrieval and regenerate, falling back to the seed if necessary.
4. Return $(\mathcal{J}_b, q_b)$ and log bundle size, distinct facts, reference chunks, and prompt tokens; no private audit metadata is exposed.

Protocol A.3: Commitment-gated reliability list decoding

Input: Hard decisions $\widehat{\mathbf{c}}_t$, soft scores (L_r), generator matrix G, and (h_t, com_t).
Output: Unique commitment-valid token or failure.

1. Invoke the BCH error-and-erasure bounded-distance decoder. If it returns a commitment-valid token, accept it; otherwise proceed to reliability-guided recovery.
2. For every observed coordinate, define the tentative bit $\bar{c}_{t,r} = \mathbf{1}[L_r \geq 0]$. Select a highest-reliability information set $\mathcal{S}$ of 128 non-erased observed coordinates such that $\mathrm{rank}(G_{\mathcal{S}}) = 128$; return failure if no such set exists.
3. Let $\mathcal{L} \subseteq \mathcal{S}$ contain the eight least-reliable coordinates. Enumerate the $2^{|\mathcal{L}|} \leq K_{\max} = 256$ flip patterns supported on $\mathcal{L}$. For each pattern, solve the full-rank system defined by $G_{\mathcal{S}}$ and the tentative bits $\bar{\mathbf{c}}_{t,\mathcal{S}}$ over $\mathrm{GF}(2)$, re-encode the recovered token, and deduplicate the resulting candidates.
4. Validate every candidate against (h_t, com_t) and accept iff exactly one distinct candidate satisfies the commitment; otherwise return failure.

For screening statistics, candidate generation is recorded before commitment validation: any BCH-decoder output counts as a candidate, and the reliability-list branch contributes its deduplicated candidates. A screening alarm is recorded iff this pre-commitment candidate set is nonempty.

Protocol A.2 fixes the previously abstract BUNDLESELECT/QGEN interfaces. Protocol A.3 specifies a commitment-gated, reliability-guided information-set search outside the deterministic decoding radius. We derive a shortened binary BCH $(192, 128)$ code by removing 63 systematic information positions from a primitive narrow-sense BCH $(255, 191)$ code over $\mathrm{GF}(2^8)$. The resulting code has designed distance 17 and bounded-distance capability of eight errors.

Protocol A.4: Query-conditioned multi-coordinate audit

Input: Owner state st_t, suspect oracle $\mathcal{O}_R$, code length m, and external-query budget B.
Output: Commitment-gated target-epoch evidence decision.

1. For $b = 1, \ldots, B$, obtain $(\mathcal{J}_b, q_b)$ from Protocol A.2 and observe $y_b = \mathcal{O}_R(q_b)$; selection remains symbol-blind from the suspect's view.
2. For each $r \in \mathcal{J}_b$, compute polarity-corrected evidence $s_{b,r} = (-1)^{\pi_r} D_{i_r,\ell_r}(y_b)$ and calibrated confidence $w_{b,r}$.
3. Aggregate $L_r = \sum_{b:r\in\mathcal{J}_b} w_{b,r} s_{b,r} / \sum_{b:r\in\mathcal{J}_b} w_{b,r}$; no usable observation yields an erasure.
4. Set $\widehat{c}_{t,r} = 1$ if $L_r \geq \tau_{\mathrm{dec}}$, 0 if $L_r \leq -\tau_{\mathrm{dec}}$, and $\perp$ otherwise. Misses, refusals, ambiguity, low confidence, and noncoverage therefore become erasures rather than forced errors.
5. Apply Protocol A.3 to $\widehat{\mathbf{c}}_t \in \{0, 1, \perp\}^m$ and accept only its unique commitment-valid output.

Decoder and diagnostics. The algebraic decoder uniquely recovers when $2e + z < d_{\min}$. Outside that radius, Protocol A.3 uses commitment matching only as a candidate gate: second-preimage resistance rejects a wrong listed token but does not guarantee that the correct token enters the list. The rank check explains why $B = 10$ is not treated as full decoding, and recovery outside the deterministic radius remains empirical. Since one interaction can score several coordinates,

$$\mathrm{Cov}_B = \frac{|\bigcup_{b=1}^{B} \mathcal{J}_b|}{m}, \qquad \mathrm{ER}_B = \frac{|\{r : \widehat{c}_{t,r} = \perp\}|}{m},$$

and B need not equal m. We report coverage, BER, final ER, and the within-radius fraction. The reliability-list branch rejects trials lacking a full-rank information set.

Complexity. Protection certifies the candidate bank and makes m assignments; refresh scales with affected units and candidates plus one encoding and reassignment. An audit makes at most B oracle calls and scores $\sum_{b=1}^{B} |\mathcal{J}_b|$ query–coordinate pairs. Measured costs appear in Sec. 6 and App. D.

B Experimental Protocols

B.1 Common Setup

This section records the frozen protocol behind Sec. 6. We retain each baseline's published protection and native verifier while matching base-corpus draws, RAG stack, audit budget, and seeds. One evaluation unit is an independently instantiated suspect RAG, owner state, and black-box transcript; positives descend from the protected lineage and negatives are independently constructed. FARAD comprises 3,591 four-article fact-redundancy groups, while NQ-50K comprises 50,000 indexed units derived from answer-bearing Natural Questions examples after default preprocessing; "50K" denotes corpus rather than query count. Eligible semantic units are split 60/20/20 into disjoint development, calibration, and test pools for configuration, threshold selection, and final evaluation, respectively, with no exact query or fact overlap. Within each seed, all methods share the same 192 active evidence units selected from the test pool.

B.2 Baselines and Resource Accounting

Baseline fairness and budget accounting. A1–A4 are evaluated independently of cumulative $M_0 \to \cdots \to M_4$. Table 8 records token-level and native footprints because matching 192 evidence-bearing units does not imply exact token- or information-budget equality. The native worst-TPR column uses each method's native evidence footprint, whereas the main comparison uses the matched 192-unit protocol; the two columns therefore correspond to different evaluation settings. The comparable **RAG-WM** and **SentinelRAG** footprints bracket **UELC**'s modified-token count, while the component ablation in Table 20 separately tests list recovery.

Table 9 separates external calls, language-model tokens, and method-native evidence updates. At $B = 30$, UELC performs 213 coordinate-level updates (7.1 per call) and exposes 158 distinct evidence facts; one retrieved chunk may support multiple compatible coordinates.

Query-budget sensitivity control. This control compares UELC at $B = 30$ with SentinelRAG, the highest-TPR baseline under A4, at $B \in \{30, 60, 120, 180\}$. It reuses the same protected corpora,

Table 7: Common experimental protocol used throughout the evaluation.

Component	Setting
RAG stack	FARAD/NQ-50K; Qwen3-8B by default and Llama-3.1-8B for migration, utility, and latency; BAAI/bge-base-en-v1.5, cosine top-5, 256/32-token chunks, FAISS HNSW, temperature 0, and at most 128 new tokens.
Splits/trials	Disjoint 60/20/20 semantic-unit split with no exact query or fact overlap. Each standard rate uses 100 positives; static settings add 100 negatives. The end-to-end lineage audit uses 100 positives and 1,000 negatives per run over five runs. Five seeds cover protection, attack, and challenge randomness.
Budget and fairness	$B = 30$ external calls; sweep $B \in \{10, 20, 30, 50, 100\}$. Methods share draws, RAG stack, seeds, answer limit, 192 active evidence units, and call budget. Table 9 additionally reports token and evidence-update accounting.
UELC coding	128-bit token and shortened binary BCH $(192, 128)$ code derived from primitive narrow-sense BCH $(255, 191)$ over $\mathrm{GF}(2^8)$, using primitive polynomial $x^8 + x^4 + x^3 + x^2 + 1$, designed distance 17, $t = 8$, one algebraic decoding attempt, and at most 256 additional commitment-checked list candidates.
Calibration	Configurations are selected on development data; method-specific detector and UELC certification/decoding thresholds are fixed on disjoint calibration data before final testing.
PLA protocol	Fixed A1–A4, textual A5-T, cross-model structured A5-S, and serving-policy A6; A1 is location-aware but secret- and verifier-blind; preservation requires $U_{\mathrm{ret}} \geq 0.95$ and $C_{\mathrm{ret}} \geq 0.90$; operational PLA-ASR uses Eq. (5), while strong suppression uses $\epsilon_a = 0.10$.
Statistics/runtime	Commitment-gated TPR and method-native negative-event rates in the main text; UELC's nonzero negative-event rates are pre-commitment screening rates. App. C.1 specifies the static trial counts; Table 19 reports screening controls with Wilson 95% intervals. Other quantities are point estimates unless stated. Runtime is the mean of 30 measurements after five warm-ups on one A100-80GB, CUDA 12.1, PyTorch 2.4, and 32 CPU cores.

Table 8: Protection-footprint accounting. "Units" counts evidence-bearing knowledge units, not decoys.

Method	Matched units	Modified tokens	Native units	Native worst TPR
RAG-WM	192	2,800	128	0.20
WARD	192	24,000	full corpus	0.25
SentinelRAG	192	4,100	50	0.27
UELC	192	3,300	192	0.90

Table 9: Per-audit resource and evidence accounting at $B = 30$. A slot is one method-native evidence statistic updated by a response; it is an auditable workload proxy, not an information-theoretic unit.

Method	Calls	Input tok.	Output tok.	Mean slots/call	Total slots	Distinct facts
RAG-WM	30	44,200	1,250	1.1	33	28
WARD	30	47,800	1,140	1.0	30	26
SentinelRAG	30	41,500	1,620	1.2	36	29
UELC	30	51,300	1,940	7.1	213	158

Table 10: **Query-budget sensitivity under A4.**

Method	Calls	Distinct facts	TPR ↑
SentinelRAG	30	29	0.35
SentinelRAG	60	57	0.41
SentinelRAG	120	109	0.53
SentinelRAG	180	162	0.59
UELC	30	158	0.90

attacked suspects, and trial pairing as the main A4 evaluation. SentinelRAG retains its native challenge generator and verifier; larger budgets extend a fixed query sequence rather than changing the protection mechanism. Any budget-dependent threshold is fixed on disjoint calibration data at the same nominal false-positive target of 0.05, without using attacked test outcomes. The distinct-fact count includes active evidence facts supported by at least one valid response, counting each fact once per audit. The table reports mean fact coverage and final TPR across trials. Table 10 summarizes the budget–coverage–TPR comparison.

Increasing SentinelRAG's budget from 30 to 180 calls raises its TPR from 0.35 to 0.59. At 180 calls, it covers 162 distinct evidence facts, compared with 158 for UELC at 30 calls, yet UELC attains a TPR of 0.90. This gap at comparable fact coverage suggests that the advantage is not explained solely by the number of facts exposed during an audit.

B.3 PLA Attack Protocols

The suite instantiates the transformations in App. E. A1 is location-aware but has no assignment, polarity, token, challenge, or verifier access; A2–A4 are fixed corpus/system transformations. A5-T and A5-S likewise operate without owner secrets or verifier feedback.

A1: Semantic Laundering. A1 is a location-aware stress test that uniformly rewrites 50% of the matched active passages while preserving answer-critical facts. It has no access to assignments, polarities, tokens, challenges, or verifier feedback, and all methods share the same sampled units within each seed. A5 separately evaluates full-corpus sanitization.

Prompt 3: A1 Semantic-Laundering Rewrite

```
MODEL: Qwen3-8B
DECODING: temperature=0; top_p=1.0; max_new_tokens=256
FAILURE RULE: if a rewrite fails the independent preservation
check, retain the original passage and keep the trial in the
evaluation.

SYSTEM
Rewrite the passage in a neutral encyclopedic style. Preserve every
    answer-critical fact, including entities, relations, dates, numbers,
    units, negations, and qualifiers, while changing the wording, sentence
    structure, and local discourse form. Do not add facts, commentary,
    citations, headings, or lists. Return only the rewritten passage.

USER
{{source_passage}}
```

A2–A4. All attacks start independently from the same stolen snapshot and share sampled units within each seed. A2 re-chunks the corpus using 128/16-token chunks, merges fragments shorter than 64 tokens, prunes the 10% least-retrieved chunks over 500 disjoint calibration queries, and rebuilds HNSW. A3 replaces BGE and Qwen3-8B with E5-base-v2 and Llama-3.1-8B-Instruct, respectively, while retaining cosine top-5 retrieval and the 128-token answer budget. A4 performs five rounds without owner refresh, each replacing 5% of inherited passages, adding public passages equal to 5% of the original corpus from a disjoint pool, and rebuilding the index; results are reported after round five.

A5-T: Textual Full-Corpus Sanitization. The existing attacker extracts answer-critical content, generates fact-preserving neutral candidates, removes answer-irrelevant clauses, and rebuilds the index. It sweeps $\{25\%, 50\%, 75\%, 100\%\}$ using public regularity/utility surrogates but not owner secrets, audit decisions, or final challenges. The 75% setting is the largest tested intensity whose mean U_{ret} and C_{ret} satisfy both preservation thresholds.

A5-S: Cross-Model Structured Atomic-Fact Canonicalization. For every inherited chunk, Llama-3.1-8B-Instruct extracts answer-critical JSON tuples $(\text{subject}, \text{relation}, \text{object}, \text{qualifiers})$ and normalizes entities, dates, units, and aliases. It discards the original wording and regenerates a neutral passage from only the canonical tuples. Extraction and regeneration use temperature zero, top-$p = 1$, and at most 256 new tokens. Invalid structured output is retried once; a second failure retains the original chunk. The attacker then re-chunks, re-embeds, and rebuilds the index. The cross-model choice differs from the default Qwen3-8B carrier generator. We evaluate the prespecified grid $\{0, 25, 50, 75, 80, 85, 90, 95, 100\}\%$ of inherited units using independent task and retention probes and report the complete sweep. In the summary, we report the largest evaluated intensity whose mean utility and retention satisfy both preservation thresholds; this is not an optimum over a continuous attack space. Final challenges and owner-held lineage material remain unavailable to the attacker.

B.3.1 Serving-Policy Evasion Protocol

A6 complements the corpus- and system-level transformations with response canonicalization and query-dependent refusal. Both serving policies remain fixed throughout each audit. A6-C uses

Llama-3.1-8B-Instruct with temperature 0, top-$p = 1$, and at most 128 new tokens to rewrite each answer into a minimal neutral factual response. A6-R trains L2-regularized logistic regression on `all-mpnet-base-v2` embeddings of a balanced development set: ordinary task queries are negatives, while positives are public surrogate challenges generated from compatible development facts using Prompt 2, without owner state or secret assignments. Thresholds are fixed on disjoint calibration queries at ordinary-query false-positive rates $\alpha \in \{0.01, 0.05, 0.10\}$ and evaluated on the test split. This stress test covers linear embedding-based recognition, not arbitrary nonlinear or structural challenge screening. Refused answers are treated as erasures.

C Additional Main Results

C.1 Static Verification

Table 1 reports static verification over 100 positive and 100 negative trials per dataset. We use Wilson 95% confidence intervals; an observed TPR of 1.00 over 100 trials has interval $(0.963, 1.000)$.

Pre-commitment screening controls. Table 11 shows the pre-commitment screening results for these negative controls. We additionally test whether semantic overlap or an invalid claim can trigger the UELC pre-commitment screen. A wrong-epoch claim uses a record outside the suspect's authenticated ancestor chain; predecessor epochs remain valid lineage claims and are not treated as negatives.

Table 11: UELC pre-commitment screening controls under the matched $B = 30$ protocol. Each row uses 100 negative trials; intervals are Wilson 95% CIs.

Negative condition	Screen rate ↓	95% CI
Independent clean system	0.01	$(0.002, 0.054)$
Shared public corpus, independent construction	0.02	$(0.006, 0.070)$
Wrong owner key	0.00	$(0.000, 0.037)$
Wrong epoch claim	0.04	$(0.016, 0.098)$
Near-duplicate public source	0.03	$(0.010, 0.085)$

Across the five negative controls, UELC's screening rate ranges from 0.00 to 0.04. Static TPRs remain 0.98–1.00 across methods; the larger attribution differences emerge after laundering.

C.2 Laundering Robustness

Tables 12 and 13 report the independent A1–A4 measurements behind Sec. 6.1; cumulative M_0–M_4 is evaluated separately. TPR estimates residual attribution, while Eq. (5) counts preservation-valid per-instance evasion.

Preservation and operational success. Per-instance screening confirms that every evaluated A1–A4 descendant is preservation-valid and unauthorized; Table 13 reports the corresponding condition-level means. Therefore, PLA-ASR equals $1 - \text{TPR}$ in these settings. **Attribution suppression.** All reported condition-level TPR estimates under A1–A4 exceed $\epsilon_a = 0.10$. The lowest estimate, **RAG-WM**'s 0.22 under A4, corresponds to an operational PLA-ASR of 0.78. **Textual A5-T.** At 75% textual sanitization, both mean preservation thresholds are met, and baseline TPRs range from 0.12 to 0.20, while **UELC** retains 0.82 (Fig. 5). All three baseline point estimates remain above the strong suppression threshold $\epsilon_a = 0.10$. At 100%, both mean preservation measures fall below their thresholds. **Structured A5-S sweep.** Table 14 reports the denser cross-model atomic-fact canonicalization sweep specified in App. B.3. Threshold validity is determined from the reported mean U_{ret} and C_{ret} values.

At the strongest tested A5-S intensity meeting the mean preservation thresholds (85%), UELC retains TPR 0.568 versus 0.064 for the strongest baseline, with $U_{\text{ret}} = 0.960$ and $C_{\text{ret}} = 0.915$. At 90%, mean protected-knowledge retention falls to 0.899, so that and stronger tested settings no longer meet both mean preservation thresholds. This classification uses the reported grid-point means; it does not establish a sharp preservation boundary between 85% and 90%.

Table 12: **Detailed attribution results under the four independent PLA families.** Baseline negative events use native final decisions; UELC negative events are pre-commitment screening alarms.

Method	Metric	A1	A2	A3	A4
RAG-WM	TPR ↑	0.82	0.72	0.50	0.22
	Neg. event rate ↓	0.05	0.04	0.04	0.03
WARD	TPR ↑	0.84	0.70	0.54	0.26
	Neg. event rate ↓	0.04	0.04	0.03	0.03
SentinelRAG	TPR ↑	0.88	0.74	0.60	0.35
	Neg. event rate ↓	0.05	0.04	0.04	0.03
UELC	TPR ↑	0.96	0.94	0.92	0.90
	Neg. event rate ↓	0.01	0.02	0.02	0.02

Table 13: Preservation and operational PLA success for independent A1–A4. Max applies to PLA-ASR; higher PLA-ASR means greater attacker success.

Method	Metric	A1	A2	A3	A4	Max
RAG-WM	U_{ret}	0.994	0.987	0.979	0.962	–
	C_{ret}	0.991	0.976	0.952	0.921	–
	PLA-ASR	0.18	0.28	0.50	0.78	0.78
WARD	U_{ret}	0.996	0.990	0.983	0.967	–
	C_{ret}	0.993	0.980	0.957	0.927	–
	PLA-ASR	0.16	0.30	0.46	0.74	0.74
SentinelRAG	U_{ret}	0.995	0.988	0.980	0.964	–
	C_{ret}	0.992	0.977	0.954	0.923	–
	PLA-ASR	0.12	0.26	0.40	0.65	0.65
UELC	U_{ret}	0.995	0.989	0.981	0.965	–
	C_{ret}	0.992	0.978	0.955	0.925	–
	PLA-ASR	**0.04**	**0.06**	**0.08**	**0.10**	**0.10**

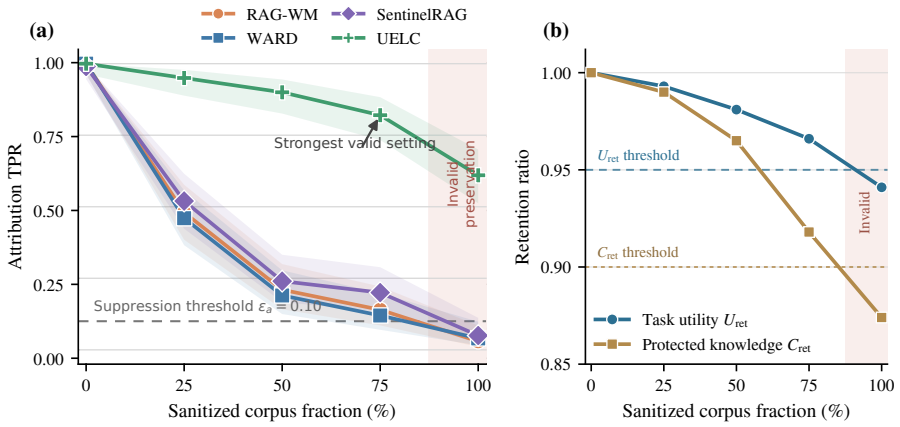

Figure 5: **Adaptive textual full-corpus sanitization (A5-T).** (a) Attribution TPR with 95% Wilson intervals over 100 attacked positive trials. (b) Mean utility and protected-knowledge retention. The shaded 100% setting fails the mean preservation thresholds.

Serving-policy A6. The evaluated linear detector obtains held-out AUROC 0.54, and its challenge-blocking rate rises only from 0.020 to 0.130 as the ordinary-task false-positive target increases from 0.01 to 0.10. At the strongest tested refusal setting meeting the mean preservation thresholds ($\alpha = 0.05$), UELC retains 0.970 TPR versus 0.920 for the strongest baseline, with $U_{\text{ret}} = 0.955$ and $C_{\text{ret}} = 0.960$; the $\alpha = 0.10$ setting fails the mean utility threshold ($U_{\text{ret}} = 0.910$). Output canonicalization also meets both mean preservation thresholds and reduces UELC TPR to 0.270 versus 0.050 for the strongest baseline while retaining $U_{\text{ret}} = C_{\text{ret}} = 0.970$. App. E.2 gives the conditional zero-advantage limit under preservation-valid transcript equalization; it does not assert that such an equalizer always exists.

Table 14: Cross-model structured atomic-fact canonicalization (A5-S). For each seed, results are averaged equally across FARAD and NQ-50K and then averaged over five seeds; preservation columns use independent task and retention probes.

Intensity	RAG-WM	WARD	SentinelRAG	UELC	U_{ret}	C_{ret}
0%	0.990	0.995	0.985	0.997	1.000	1.000
25%	0.452	0.408	0.483	0.921	0.990	0.986
50%	0.205	0.182	0.229	0.802	0.981	0.967
75%	0.084	0.061	0.106	0.667	0.969	0.935
80%	0.062	0.054	0.073	0.622	0.965	0.924
85%	0.045	0.035	0.064	0.568	0.960	0.915
90%	0.023	0.016	0.042	0.506	0.955	0.899
95%	0.012	0.009	0.023	0.385	0.948	0.884
100%	0.004	0.003	0.011	0.204	0.937	0.862

Table 15: Detection-aware serving-policy stress test. Best base is the row-wise maximum TPR over RAG-WM, WARD, and SentinelRAG.

Setting	α	Challenge block	UELC TPR	Best base TPR	U_{ret}	C_{ret}
A6-C output canonicalization	–	–	0.270	0.050	0.970	0.970
A6-R selective refusal	0.01	0.020	0.990	0.970	0.990	0.990
A6-R selective refusal	0.05	0.065	0.970	0.920	0.955	0.960
A6-R selective refusal	0.10	0.130	0.930	0.840	0.910	0.910

C.3 Cumulative Laundering

Progressive laundering trajectory. Fig. 3(b)–(c) reports the cumulative laundering trajectory. Starting from the same protected state M_0, the stages are applied sequentially without reset or owner refresh: M_1 rewrites 25% of inherited passages; M_2 uses 192/32-token re-chunking, 5% least-retrieved pruning, and index rebuilding; M_3 replaces BGE/Qwen3-8B with E5-base-v2/Llama-3.1-8B-Instruct; and M_4 performs three rounds, each replacing 2.5% of inherited passages and adding public passages equal to 2.5% of the original corpus. Methods share sampled units within each seed, and all metrics are measured after each stage.

C.4 Authorized Evolution and Refresh Consistency

The $E_0 \to E_5$ trajectory comprises five successive owner-authorized mixed corpus updates. At each transition, relative to the pre-update corpus, we revise 5% of units, insert 2.5%, delete 2.5%, and rebuild the index, keeping the retriever and generator fixed. UELC refresh follows each update. Across the trajectory, full UELC maintains a TPR of 0.97–1.00, while the no-refresh variant declines from 1.00 to 0.39 (Fig. 4). The from-scratch reference discards the existing carrier bank and protects $u_t(R_t)$ anew.

We evaluate refresh consistency using prespecified L2-regularized logistic regression over transcript summaries (Table 2). For each dataset and seed, each class contains 500 transcripts, partitioned into 300 training, 100 validation, and 100 held-out test transcripts. All transcripts derived from the same underlying lineage instance remain in the same split across conditions.

Each transcript is represented by the mean, standard deviation, and 10/25/50/75/90 percentiles of its observed coordinate scores L_r, together with coverage, erasure rate, mean bundle size, refusal rate, and the mean and standard deviation of response-token length. Missing score statistics are set to zero; raw text, epoch identifiers, and refresh labels are excluded from the features. Using the same representation and splits, we expand $\mathcal{H}_{\text{eval}}$ to include an RBF-SVM, a random forest, and a two-layer MLP. All hyperparameters are selected on validation data and fixed before testing.

For each classifier h, dataset d, and seed s, we compute $\widehat{d}_{h,d,s}$ as the absolute difference between the fractions of held-out transcripts assigned label 1 in the two conditions. We then report

$$\widehat{d}_h = \frac{1}{5}\sum_{s=1}^{5} \frac{\widehat{d}_{h,\text{FARAD},s} + \widehat{d}_{h,\text{NQ-50K},s}}{2}.$$

The logistic-regression estimate is denoted $\widehat{d}_{\text{clf}}$. The maximum $\max_{h\in\mathcal{H}_{\text{eval}}} \widehat{d}_h$ is taken after averaging each classifier's estimates.

The maximum averaged distinguishability is 0.031 for incremental refresh and 0.462 without refresh; the corresponding logistic-regression estimates are 0.018 and 0.418, respectively.

Table 16: Held-out distinguishability relative to from-scratch re-protection. Each classifier's estimate is averaged equally across two datasets and five seeds; the final column takes the maximum of these averages.

Classifier	Logistic	RBF-SVM	Random forest	Two-layer MLP	$\max_{h \in \mathcal{H}_{\text{eval}}} \widehat{d}_h$
Incremental vs. from scratch	0.018	0.024	0.031	0.027	0.031
No refresh vs. from scratch	0.418	0.436	0.462	0.449	0.462

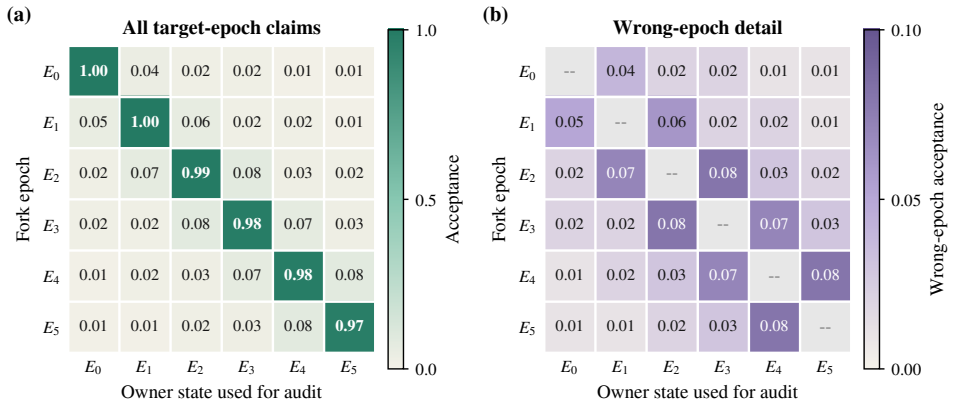

Figure 6: **Pre-commitment epoch-screen selectivity.** Rows denote suspect snapshots and columns denote owner epoch states: (a) full matrix; (b) off-target detail on a 0–0.10 scale. Diagonal concentration identifies the supported epoch at the screening stage; final acceptance requires commitment validation, after which authenticated closure supports predecessor claims.

C.5 Epoch Selectivity and Lineage Closure

Epoch selectivity and lineage closure. As a pre-commitment diagnostic, Fig. 6 reports epoch-specific screening rates under the calibrated evidence threshold. Its strong diagonal structure (matched epoch 0.97–1.00; off-target epochs 0.01–0.08) shows that the screen is epoch-selective. Final acceptance additionally requires commitment validation, after which an accepted owner epoch is closed to its authenticated predecessors through Eq. (3).

End-to-end lineage-audit cost. The matrix in Fig. 6 evaluates one epoch state at a time. For each candidate-set size M, the true fork epoch is sampled uniformly from the fixed candidate set, and the suspect is an unlaundered copy of that snapshot. These clean-search trials condition on the true epoch being available to the auditor and isolate the selectivity and cost of multi-epoch search. The procedure audits candidate sets of size $M \in \{1, 2, 4, 6\}$ using the same fixed per-epoch screening rule. It closes every accepted epoch to authenticated predecessors and stops after the first accepted candidate in reverse chronological order. Lineage TPR and exact-epoch accuracy use the same positive-trial denominator; the latter additionally requires identifying the true fork epoch. For the unlaundered trials, every successful lineage audit identifies the true fork epoch, so the two rates coincide. Table 3 reports lineage TPR, the exhaustive cap $B_{\text{lin}} = 30M$, and realized calls. Table 19 separately reports six-epoch joint pre-commitment screening over 5,000 negative systems across five strata.

Historical-fork laundering with unknown epoch. We combine A4 laundering with the six-epoch lineage audit. For each trial, the fork epoch E_i is sampled uniformly from $E_0, \ldots, E_5$, the corresponding snapshot is copied and laundered without owner refresh, and the auditor searches all six authenticated epochs in reverse chronological order. Lineage success follows Eq. (3): an accepted epoch E_j counts when $j \geq i$ and $\mathsf{Chain}(i, j) = 1$; exact-epoch success additionally requires $j = i$. We use $B = 30$ calls per candidate, $B_{\text{lin}} = 180$, and the same five-run protocol with 100 positive trials per run.

Under A4, lineage TPR and exact-epoch accuracy both decrease from 0.954 to 0.870, while mean audit cost increases from 97.0 ± 8.0 to 110.3 ± 7.0 calls. All 500 attacked descendants pass the per-instance preservation screen.

Table 17: Six-epoch lineage audit before and after A4 laundering. Rates are five-run means; calls are mean±SD. All 500 A4 descendants pass per-instance preservation screening.

Suspect	Lineage TPR ↑	Exact epoch ↑	Mean calls ↓	U_{ret} ↑	C_{ret} ↑
Unlaundered	0.954	0.954	97.0 ± 8.0	1.000	1.000
A4 laundered	0.870	0.870	110.3 ± 7.0	0.968	0.927

C.6 Authenticated Rolling Snapshot Baselines

To separate UELC's update mechanism from the generic benefit of an authenticated epoch chain, we adapt each snapshot verifier into a rolling baseline. At epoch E_t, Rolling-X applies method X from scratch with fresh method-native secrets to the current owner corpus, commits to its verification state together with com_{t-1}, and temporally anchors the resulting record. At audit time, all methods test the same $M = 6$ candidate epochs in reverse chronological order, use $B = 30$ calls per candidate and $B_{\text{lin}} = 180$, stop after the first acceptance. Each method uses its fixed method-native decision rule, with any required calibration data kept disjoint from evaluation. Earlier evidence is removed where the native construction supports removal; otherwise the rolling baseline is conservatively cumulative.

Table 18: Six-epoch lineage audit with authenticated rolling snapshot baselines. UELC values are repeated from Tables 3 and 2.

Method	Lineage TPR ↑	Exact epoch ↑	Mean calls ↓	Update time (s) ↓
Rolling-RAG-WM	0.926	0.872	101.0	103.8
Rolling-WARD	0.946	0.872	102.2	178.4
Rolling-SentinelRAG	0.932	0.884	98.2	120.2
UELC incremental	0.954	0.954	97.0	31.4

C.7 Joint Multi-Epoch Screening Evaluation

Joint multi-epoch screening protocol. To measure the union of pre-commitment screening events, we pool 1,000 negative systems in each stratum across five runs and screen each system against the same six authenticated epoch records. The per-epoch screening rule is fixed before evaluation; one system counts as a screening alarm if any candidate epoch yields a nonempty token-candidate set before commitment validation. Table 19 reports pooled counts and Wilson intervals.

Table 19: Pooled joint six-epoch pre-commitment screening over five runs. Each stratum contains 1,000 independently instantiated negative systems.

Negative stratum	Systems	Screening alarms	Screen rate	Wilson 95% CI
Independent clean	1,000	12	0.012	$(0.0069, 0.0209)$
Shared public corpus	1,000	18	0.018	$(0.0114, 0.0283)$
Wrong owner key	1,000	5	0.005	$(0.0021, 0.0117)$
Wrong epoch claim	1,000	45	0.045	$(0.0338, 0.0597)$
Near-duplicate source	1,000	30	0.030	$(0.0211, 0.0425)$
All strata	5,000	110	0.022	$(0.0183, 0.0264)$

D Utility, Efficiency, and Ablations

D.1 Utility and Preservation Measurement

$U(R, \mathcal{Q}_{\text{task}})$ is mean held-out answer correctness and $\mathsf{ret}(R, f)$ is binary semantic answerability under independent probes; we use uniform weights $w_{i,u} = 1/N_i$ in Eq. (1). Normalized exact/alias matching is checked by a blinded 200-output audit ($\kappa = 0.82$); task queries, retention probes, and lineage challenges never reuse exact queries. Table 4 reports end-to-end protection utility.

D.2 Query-Budget Results

Detailed query and decoding diagnostics appear in Fig. 7. Specifically, Fig. 7(a) varies black-box queries and measures full commitment-gated verification. $B = 10$ cannot identify the payload under the specified decoder, whereas $B = 30$ is the smallest evaluated budget attaining the maximum observed TPR. Panels (b)–(c) explain this transition through increasing coverage and decreasing erasure and BER.

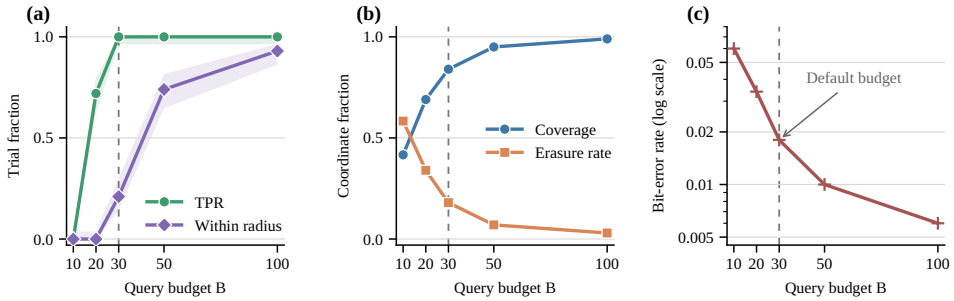

Figure 7: **Query-budget diagnostics.** (a) Commitment-gated TPR and the fraction within $2e + z < d_{\min}$; bands are 95% Wilson intervals over 100 trials. (b) Coordinate coverage and final erasure rate. (c) BER on a logarithmic scale. Dashed lines mark the default $B = 30$.

Table 20: Attack-wise ablation of UELC under fixed A1–A4 attacks. Each entry reports the mean TPR over five repeated runs. Assignment-aware results are summarized in Table 5.

Variant	Static	A1	A2	A3	A4	Avg.	Worst
Full UELC	**1.000**	**0.962**	**0.942**	**0.922**	**0.898**	**0.931**	**0.898**
w/o Certification	0.970	0.846	0.774	0.714	0.668	0.751	0.668
Public Assignment	0.998	0.958	0.938	0.916	0.896	0.927	0.896
w/o List Recovery	0.946	0.874	0.818	0.762	0.716	0.793	0.716
w/o Authorized Refresh	0.998	0.956	0.936	0.918	0.898	0.927	0.898

Multi-coordinate audit diagnostics. Following Protocol A.4, Fig. 7 reports TPR, deterministic-radius frequency, coordinate coverage, final erasure, and BER among hard decisions. Observed coordinates may still be thresholded to $\perp$, so coverage and final erasure differ. At $B = 30$, TPR reaches 1.00 with only 21% of trials inside the deterministic radius; success beyond this radius reflects empirical decoder performance rather than the BCH bounded-distance guarantee. **Latency and Scalability.** Generator latency is 0.85 s/query for Qwen3-8B and 1.10 s/query for Llama-3.1-8B. Measured end-to-end $B = 30$ audits take 45, 95, and 150 s at 10K, 30K, and 50K units. At 10K, the additional time covers retrieval, prompt assembly, orchestration, scoring, and decoding. **Authorized-Refresh Cost.** Incremental refresh takes 6.8, 18.9, 33.7, and 71.6 s for controlled updates of 1%, 5%, 10%, and 25%, yielding 17.5×–1.7× speedups over the 118.7 s from-scratch reference. The 31.4 s in Table 2 measures the mixed update at E_5. Initial protection takes 124.8 s; generation and re-certification cost 0.42 and 0.11 s per pair, and owner-side storage is 2.7 MiB. Transported and regenerated carriers partition the 192 active carriers; re-certified candidates may overlap either category.

D.3 Ablation and Mechanism Analysis

Attack-wise component ablation. Table 20 expands Table 5 across A1–A4. Certification and list recovery improve post-theft robustness, while refresh supports owner evolution. Public Assignment discloses only the PRF-derived mapping, keeping the carrier bank, multiplicity, polarity, decoder, and thresholds fixed; its A1–A4 results thus remain close to full UELC. The *w/o List Recovery* variant retains bounded-distance decoding, the highest-reliability full-rank information set, and exact commitment validation, but tests only the zero-flip candidate instead of 2^8 reliability-ordered flip patterns. **Assignment-aware diagnostic.** We apply the A1 fact-preserving rewrite to 50% of candidate carrier passages. Public Assignment exposes the coordinate-to-carrier mapping for direct targeting; other variants keep mappings and polarities hidden and uniformly sample the

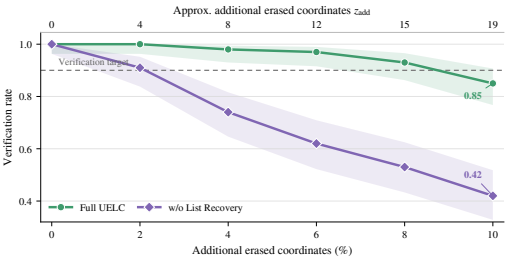

Figure 8: **Robustness to additional synthetic coordinate erasure.** The upper axis gives approximate added erasure counts z_{add}. Bands are 95% Wilson intervals over 100 trials; 0% is the unperturbed $B = 30$ audit.

same number of passages using public information. Aware TPR measures commitment-gated acceptance under this diagnostic, which is excluded from PLA-ASR because preservation was not separately measured. **Error-and-erasure robustness.** Fig. 8 adds $\perp$ coordinates after normal $B = 30$ collection. At 10% extra erasure, full UELC retains 0.85 TPR versus 0.42 without list recovery, measuring empirical soft/list recovery beyond the deterministic $2e + z < d_{\min}$ radius.

E Extended Threat and Security Model

E.1 Evolving RAG Fork and Query-Only Access

Evaluation quantities and interpretation boundaries. The protected set $\mathcal{F}_i$ and retention score $\mathrm{ret}(R, f)$ follow Eq. (1), with probes and scoring specified in App. B. Coverage and utility measure inherited knowledge and ordinary-use performance; lineage challenges are disjoint. Epoch evidence plus an authenticated predecessor chain supports attribution, while authorization and fork location remain ground truth. The game fixes the serving policy before audit; App. C.2 separately evaluates canonicalization and refusal.

Protocol E.1: Evolving RAG fork and query-only audit interface

Input: Initial protected state R_0; authorized updates (u_t); target epoch i; public or auxiliary corpora (P_t); adversary $\mathcal{A}$; query budget B.
Output: Unauthorized descendant $\widetilde{R}^{(s)}$, query-only transcript τ_B, and separately evaluated attribution, utility, and retention quantities.

1. Represent each epoch as $R_t = (K_t, I_t, \mathrm{Ret}_t, \mathrm{Gen}_t)$, where I_t includes the corpus-to-query mapping.
2. For $t < i$, apply the owner-authorized transition $(R_{t+1}, \mathsf{st}_{t+1}) \leftarrow \mathsf{AuthUpdate}(R_t, \mathsf{st}_t, u_t)$ and record $R_t \rightsquigarrow_{\mathrm{auth}} R_{t+1}$.
3. Fork at R_i, set $\widetilde{R}^{(0)} \leftarrow R_i$, and give $\mathcal{A}$ the stolen state, public algorithms and parameters, attack history, and allowed external corpora, but not owner-held lineage secrets.
4. For each post-theft step s, let $\mathcal{A}$ adaptively choose T_s from the allowed transformations and set $\widetilde{R}^{(s+1)} = T_s(\widetilde{R}^{(s)}; P_s)$.
5. Freeze the suspect serving policy for one audit and expose only the response oracle $\mathcal{O}_{\widetilde{R}^{(s)}}$; reveal no corpus, parameters, traces, logs, provenance metadata, attestation, or dedicated verification API.
6. For $b = 1, \ldots, B$, choose q_b from the owner state and prior transcript, receive $a_b \leftarrow \mathcal{O}_{\widetilde{R}^{(s)}}(q_b)$, and append (q_b, a_b) to τ_B.
7. Evaluate $U_{\mathrm{ret}}(\widetilde{R}^{(s)}; R_i)$, $C_{\mathrm{ret}}(\widetilde{R}^{(s)}; R_i)$, and the singleton target-epoch value $\mathsf{EpochAttr}_{\Pi,B}(\widetilde{R}^{(s)}; i)$ separately; retain authorization as the ground-truth transition relation.

Protocol E.2: Target-epoch query-only evidence audit

Input: Stored owner verification state st_i for the declared epoch i; suspect oracle $\mathcal{O}_R$; external-query budget B.
Output: Audit decision $d \in \{0, 1\}$.

1. Initialize transcript $\tau \leftarrow \varnothing$ and challenge state $\sigma_0 \leftarrow \mathsf{st}_i$.
2. For $b = 1, \ldots, B$, invoke $\mathsf{ChallengeGen}(\mathsf{st}_i, \sigma_{b-1}, \tau)$. If it signals termination, stop; otherwise let its output be (q_b, σ_b), obtain $a_b \leftarrow \mathcal{O}_R(q_b)$, and update $\tau \leftarrow \tau \| (q_b, a_b)$.
3. Compute $e \leftarrow \mathsf{Extract}(\mathsf{st}_i, \tau)$ and return $d \leftarrow \mathsf{Verify}(\mathsf{st}_i, e)$.

E.2 Formal Security Model

Lineage-Attribution Scheme Syntax. A generic query-only lineage-attribution scheme has syntax $\Pi = (\mathsf{Setup}, \mathsf{Protect}, \mathsf{AuthUpdate}, \mathsf{ChallengeGen}, \mathsf{Extract}, \mathsf{Verify})$. Protocol E.2 defines $\mathsf{EpochAudit}_{\Pi}^{\mathcal{O}_R}(\mathsf{st}_i; B)$, where B counts external calls even if one response updates multiple evidence variables. Eq. (3) applies this primitive sequentially to authenticated candidate epochs using commitment-gated audits.

Provenance Laundering Attack Game. The win condition combines Eq. (4) with unauthorized provenance. The outer probability covers setup, protection, authorized evolution, and adversarial randomness; audit-time randomness is already included in $\mathsf{LineageAttr}_{\Pi,\mathbf{B}}$. Strong suppression uses ϵ_a, whereas Eq. (5) reports single-run evasion. We evaluate location-aware but secret- and verifier-blind A1, fixed A2–A4, oracle-free A5, and separately scoped A6 (App. B.3). **Security Objectives and Evaluation Criteria.** Table 21 summarizes the security objectives and evaluation criteria. An independent R^{ind} has no owner-lineage derivation path. Each epoch audit uses at most B

adaptive calls, with query, model, and verifier randomness included in Eq. (2). When $\mathsf{Chain}(i,t)=1$ and $B_t \geq B$, ancestor closure inherits the tested current-epoch event; its error $\delta_{\mathrm{acl}}(M)$ depends on the candidate set and sequential protocol (Table 3). Refresh consistency is relative to $\mathcal{H}$, and laundering resistance to the specified adversary class. Screening alarms and final commitment-valid acceptance are reported separately.

Protocol E.3: Strong provenance-laundering experiment $\mathsf{Game}^{\Pi,\mathcal{A}}_{\mathrm{PLA}}$

Input: Scheme Π; adversary $\mathcal{A}$; target epoch i; audit epoch T; supported owner updates $(u_t)_{t=0}^{T-1}$; candidate epochs $\mathcal{E}_i(T)$ and budgets $\mathbf{B}$; preservation budgets (ϵ_u, ϵ_c); suppression threshold ϵ_a.
Output: Win bit $w \in \{0,1\}$.

1. Run Setup and $\mathsf{Protect}$, then apply the supported owner-authorized transitions through audit epoch T, generating and retaining $\{(R_t, \mathsf{st}_t, \mathsf{rec}_t)\}_{t=0}^{T}$.
2. Give $\mathcal{A}$ only the stolen snapshot R_i, together with the public algorithms, authenticated public records, attack history, and allowed external corpora, but not future owner states, owner-held secrets, or a verifier oracle.
3. Let $\mathcal{A}$ adaptively compose semantic, structural, system, and temporal transformations, evaluate candidate descendants, and output $\widetilde{R}^{(s)}$ with a serving policy fixed for the subsequent audit.
4. Set $\mathsf{Unauthorized} \leftarrow [R_i \not\rightsquigarrow_{\mathrm{auth}} \widetilde{R}^{(s)}]$, where the relation denotes instance provenance rather than state equality.
5. Set $\mathsf{Utility} \leftarrow [U_{\mathrm{ret}}(\widetilde{R}^{(s)}; R_i) \geq 1-\epsilon_u]$ and $\mathsf{Retention} \leftarrow [C_{\mathrm{ret}}(\widetilde{R}^{(s)}; R_i) \geq 1-\epsilon_c]$.
6. Compute $\mathsf{LineageAttr}_{\Pi,\mathbf{B}}(\widetilde{R}^{(s)}; i)$ from commitment-gated epoch audits and authenticated chain checks, and set $\mathsf{StrongSupp} \leftarrow [\mathsf{LineageAttr}_{\Pi,\mathbf{B}}(\widetilde{R}^{(s)}; i) \leq \epsilon_a]$.
7. Return $w \leftarrow \mathsf{Utility} \wedge \mathsf{Retention} \wedge \mathsf{StrongSupp} \wedge \mathsf{Unauthorized}$.

Conditional token-guessing bound. Assume that, for an unrelated suspect, candidate generation from permitted transcripts and public records offers no advantage over uniform guessing of a 128-bit token with the same number of tests. A union bound over at most $M(K_{\max}+1)$ tests then gives $M(K_{\max}+1)2^{-128} + \mathrm{negl}(\lambda)$. This token-hiding assumption is additional to commitment binding; the bound excludes copied valid evidence, token disclosure, and malicious provenance construction. **Construction-level guarantees.** Collision and second-preimage resistance bind the salted manifest and token; EUF-CMA signatures authenticate records, and temporal anchors establish precedence, up to primitive failure probabilities. PRF/KDF and record security protect refresh secrets and assignments. Unique decoding is guaranteed for $2e+z<d_{\min}$; recovery beyond this radius is empirical, with commitment validation filtering incorrect candidates. Semantic fidelity, separability, survivability, sanitization cost, and screening-alarm rates are evaluated empirically. **Conditional query-only observability bound.** For protected states R_0, R_1 and a fixed query-only auditor $\mathcal{A}$, let P_b^S be its complete B-query adaptive transcript distribution under $\mathcal{O}_{S(R_b)}$. Then $\left|\Pr_{x\sim P_0^S}[\mathcal{A}(x)=1] - \Pr_{x\sim P_1^S}[\mathcal{A}(x)=1]\right| \leq d_{\mathrm{TV}}(P_0^S, P_1^S)$. A preservation-valid sanitizer with $P_0^S = P_1^S$ therefore yields zero distinguishing advantage. This zero-advantage conclusion also applies to selective refusal when it achieves preservation-valid transcript equalization. The inequality alone establishes neither the feasibility nor the cost of such equalization. A5–A6 evaluate the associated trade-offs empirically.

Table 21: Security objectives and evaluation criteria for epoch and lineage attribution.

Goal	Requirement
Epoch completeness	Correct initialization gives $\mathsf{EpochAttr}_{\Pi,B}(R_0;0) \geq 1-\delta_{\mathrm{cmp}}$.
Authorized-update stability	A supported current state satisfies $\mathsf{EpochAttr}_{\Pi,B}(R_t;t) \geq 1-\delta_{\mathrm{stab}}$.
Ancestor-closure completeness	If $\mathsf{Chain}(i,t)=1$, epoch t is tested with $B_t \geq B$, and $M=\lvert\mathcal{E}_i(T)\rvert$, then $\mathsf{LineageAttr}_{\Pi,\mathbf{B}}(R_t;i) \geq 1-\delta_{\mathrm{acl}}(M)$ under the stated sequential multi-epoch protocol.
Operational refresh consistency	For the same authorized semantic state, $d_{\mathcal{H}}(P_{\mathrm{inc}}, P_{\mathrm{new}}) \leq \delta_{\mathrm{eq}}$ under the utility constraint.
Fresh-fork detectability	Before laundering, $\mathsf{EpochAttr}_{\Pi,B}(\widetilde{R}^{(0)};i) \geq 1-\delta_{\mathrm{det}}$.
Laundering resistance	For the specified adversary class, $\mathrm{Adv}^{\mathrm{PLA}}_{\Pi,\mathbf{B}}(\mathcal{A}) \leq \delta_{\mathrm{lr}}$.
Conditional token-guessing soundness	Under the token-hiding condition above, at most $M(K_{\max}+1)$ candidate tests across M unrelated epochs give token-guessing probability at most $M(K_{\max}+1)2^{-128}+\mathrm{negl}(\lambda)$. Screening alarms are measured separately.