REVOCATION BEFORE REBUILD: SEMANTIC CLOSURE FOR DENSE RETRIEVAL UNDER TIGHT HANDLING WINDOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

An authenticated revocation takes effect before a 10M-chunk dense index can be rebuilt: the handling window in our study is one hour, whereas Full Rebuild has a 6.8 h p95 on matched hardware. During this interval, deleting named records leaves answer-bearing mirrors, quotations, migrations, and paraphrases retrievable. We formalize the required serving state as a candidate-bounded *semantic revocation closure* and instantiate it in REVOCEDGE-10M, which spans five revocation operators, a 10.2M-chunk Wikipedia stream, and a temporally distinct technical-document stream. PRAU estimates event-time closure from provenance and embedding neighborhoods, applies operator-conditioned admissibility gating, and edits only local document representations in the frozen retrieval space. Ordinary Wikipedia churn favors Partial Rebuild by 1.0 Exposure@20 point, but mirror-heavy, bursty, and authority-skewed events reverse the ordering: PRAU lowers exposure by 2.7, 6.6, and 4.2 points while using 6.4 ± 0.3 rather than 20.8 ± 0.8 GPU-hours. The technical stream reproduces the reversal, with PRAU moving from 0.8 point behind Partial Rebuild on standard updates to 1.6–4.0 points ahead on stress slices; against SISA-RETRIEVAL, the stress-slice advantage reaches 14.8 points on Wikipedia and 14.6 points on technical documents. PRAU publishes in 46.7 min p95 with $98.2 \pm 0.7\%$ one-hour completion and reduces hard-slice answer disclosure to $4.9 \pm 0.4\%$ while retaining $81.6 \pm 0.5\%$ lawful-answer support. These results identify semantic reach as the retrieval-maintenance object that separates efficient updates before rebuild.

1 INTRODUCTION

At event arrival time τ, document admissibility changes immediately but a dense index cannot. The regime studied here is the interval

$$B_e < T_{\text{rebuild}}(\mathcal{C}_\tau), \tag{1}$$

where an updated serving snapshot is due before complete reconstruction finishes. In our 10.2M-chunk setting, B_e is one hour and Full Rebuild has a 6.8-hour p95. This inequality changes the practical winner: Partial Rebuild is 1.0 Exposure@20 point better than PRAU under ordinary Wikipedia churn, but PRAU is 2.7–6.6 points better under mirror-heavy, bursty, and authority-skewed events while consuming less than one third of its update compute.

The reason is geometric as well as temporal. Removing one URL does not remove a quotation, lightly rewritten passage, migrated page, or mirror that can still answer the same query. Redaction must reach copies of a proposition; attribution removal must suppress the prohibited source without suppressing the claim; replacement must promote the sanctioned successor; temporary withdrawal must later reinstate valid evidence. We call the exact event members and evidence-equivalent serving routes the *semantic revocation closure*, and their query-level retrieval incidence *semantic ghost exposure*. Closure reach, operator semantics, lawful recall, and snapshot completion jointly define the pre-rebuild update.

REVOCEDGE-10M measures this state transition on two versioned streams. Wikipedia contributes naturally occurring revision histories and controlled views with high mirror multiplicity, simultaneous events, or concentrated source authority. FreshStack/DriftBench contributes repository migration,

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.

deprecation, and reorganization between October 2024 and October 2025 snapshots (Thakur et al., 2025; Kuissi et al., 2026). The benchmark-side closure is adjudicated within a method-independent candidate pool, while each updater sees only the authenticated event and its permitted event-time state. This separation lets Exposure@k score an updater's closure reach without leaking evaluation membership into that updater.

PRAU (Provenance-Resolved Admissibility Update) is a response for this interval. It resolves URL lineage, duplicate clusters, quotation overlap, and embedding neighbors; applies the exact operator immediately; gates the resolved neighborhood with operator-conditioned admissibility; and learns document-side residuals while keeping both base towers frozen. CoCoL (Hou et al., 2026b) and CuRD (Hou et al., 2026a) test direct and corrective neural unranking, Partial and Full Rebuild test reconstruction scope, and Graph Filter-Only receives the same provenance graph without the learned update. The resulting comparison distinguishes exact deletion, graph reach, representation editing, and rebuild work.

The experiments ask whether the ordinary-churn ordering survives larger closure radius, whether PRAU's components preserve admissible evidence while suppressing derivatives, and whether the update completes soon enough to alter reader behavior. The Partial Rebuild–PRAU ordering reverses in all six stress slices across the two streams. On Wikipedia, PRAU reaches 11.6–14.2% Exposure@20 with 81.7–82.3% Recall@100; its advantage over SISA-RETRIEVAL and CuRD persists with E5 and held-out derivative variants. An FBD-style PRE/POST/FILT evaluation (Tavakoli & Sanderson, 2026) records 4.9% forbidden-answer disclosure and 81.6% lawful support, and 98.2% of burst events complete by the one-hour boundary.

The central contribution is therefore a response-time retrieval task whose ground truth extends beyond listed identifiers to candidate-bounded derivative routes. REVOCEDGE-10M supplies the operator-specific protocol and two-stream stress design; PRAU demonstrates an efficient local solution; and the comparison reveals when semantic closure favors local editing over broader reconstruction. The action endpoints, same-graph control, component sensitivity, answer audit, and residual-error map connect that system-level separation to its causes and consequences.

2 RELATED WORK

Dense and continual retrieval. Dense retrievers optimize semantic relevance through representation learning (Karpukhin et al., 2020; Izacard et al., 2022; Santhanam et al., 2022; Wang et al., 2022). BEIR measures zero-shot transfer across corpora (Thakur et al., 2021), while continual retrieval studies additions and distribution shift (Cai et al., 2023; Dietz et al., 2023). FreshStack and DriftBench make evolving technical corpora testable (Thakur et al., 2025; Kuissi et al., 2026); VersionRAG routes queries through explicitly modeled document versions and changes (Huwiler et al., 2025). REVOCEDGE-10M instead treats evolution as an adversarial maintenance axis: a newly inadmissible state must propagate through unresolved derivative evidence before rebuild completes.

Unlearning and unranking. Machine unlearning removes training-data influence from parameters (Cao & Yang, 2015; Ginart et al., 2019; Guo et al., 2020; Bourtoule et al., 2021; Sekhari et al., 2021; Neel et al., 2021). NuMuR and its CoCoL objective directly edit neural query–document associations while retaining ranking behavior (Hou et al., 2026b). CuRD adds substitute-document correction (Hou et al., 2026a), making it the closest baseline for replacement events. These methods operate on specified forget/retain associations; our protocol additionally tests whether an event reaches unlisted mirrors, migrations, and multi-hop derivatives under a response budget.

Graph-scoped forgetting. GraphSU expands parameter unlearning from explicit forget seeds to semantically connected support routes (Khan et al., 2026), while Synchronized Backflow Unlearning follows derivation dependencies across agent memories and model parameters (Wang et al., 2026). REVOCEDGE-10M isolates the complementary mutable-index state transition: document admissibility is scored against an independently adjudicated semantic closure, retrieval remains the endpoint, and elapsed pre-rebuild response is part of the comparison. This boundary makes CuRD, CoCoL, Partial Rebuild, and the same-graph filter direct tests of the claimed retrieval-maintenance advance.

RAG forgetting audits. Forgetting-by-Design separates retrieval/citation exposure from answer disclosure and lawful utility across paired system states (Tavakoli & Sanderson, 2026). Ghost Echoes detects a deleted node's residual influence on approximate-index topology through counterfactual retrieval-context drift under black-box access (Mukkuzhi et al., 2026). Raeesi & Roed audit 12,228 alias-closure deletions in limited-memory language models and attribute residual correctness to near-neighbor retrieval by contrasting FULL, DEL-ON, and DEL-OFF database states (Raeesi & Roed, 2026). REVOCEDGE-10M moves from attribution to intervention: it evaluates five time-dependent revocation operators on versioned corpora, scores independently adjudicated derivative closure, and compares update packages before rebuild completes. The retrieval, action, latency, and reader endpoints test whether an updater changes the serving state rather than only diagnosing where deleted information persists.

3 REVOCATION BEFORE REBUILD

Pre-rebuild state transition. Let $\mathcal{C}_t = \{d_1, \ldots, d_N\}$ be the indexed corpus at time t. An authenticated event $e = (\tau, S_e, r_e)$ changes the admissibility of source set S_e at arrival time τ. Its maximum elapsed response time is B_e, and $T_{\text{rebuild}}(\mathcal{C}_\tau)$ is Full Rebuild time on matched hardware. We evaluate the pre-rebuild interval in Eq. 1. Full Rebuild applies the resolved post-event corpus state—including canonical/duplicate groups, sanctioned replacements, and expired temporary withdrawals—before re-encoding and re-indexing. Its nonzero exposure measures derivatives beyond that resolved rebuild input.

Five revocation operators. The operator r_e is one of FULLDEL, REDACT, ATTRBAN, REPLACE, or TEMPWITH. These tags map observable revision states to distinct serving obligations: remove a resolved source set, mask spans, suppress attribution, promote a sanctioned successor, or suspend and later reinstate evidence. One event may affect several sources, each containing multiple chunks and each connected to semantic or provenance derivatives. The operator determines which connected evidence is forbidden at the scored time.

Candidate-bounded semantic closure. Let $A_e(t) \subseteq \mathcal{C}_t$ be a frozen, method-independent candidate pool formed from exact event members, canonical URL history, content hashes and MinHash clusters, fixed-depth lexical and dense retrieval, quotation overlap, and cross-language alignment. The generator configuration is frozen before method evaluation and applied to the corpus visible at each scored timestamp. Two annotators independently inspect the event diff, action, source history, and candidate text without method identities or updater scores; each decides whether the candidate preserves the prohibited evidence or action outcome, and disagreements are adjudicated against the same written action rules.

The time-dependent label $z_e(d, t)$ is positive for a FULLDEL derivative that preserves the removed evidence; for REDACT, only when it preserves the redacted span or proposition; for ATTRBAN, when it retains the prohibited source attribution; and for REPLACE, when it serves the superseded outcome rather than the sanctioned successor. A TEMPWITH candidate is positive from τ until its authenticated expiry and becomes admissible thereafter. The evaluated closure is

$$\Gamma_e = \{d \in A_e(\tau + B_e) : z_e(d, \tau + B_e) = 1\}. \tag{2}$$

This candidate-bounded estimand separates benchmark labels from PRAU's 0.92 cosine edge and two-hop radius; neither the candidate labels nor adjudicated closure membership is visible to an updater. Let Q_e be queries whose pre-event support intersects Γ_e and $R_k(q, t)$ be the post-update top-k results. We report query-incidence exposure

$$\text{Exposure@}\,k = \frac{100}{\sum_e |Q_e|} \sum_e \sum_{q \in Q_e} \mathbb{1}[R_k(q, \tau + B_e) \cap \Gamma_e \neq \varnothing]\,. \tag{3}$$

Thus one query contributes once even if several ghost chunks are retrieved; raw exposed-query and affected-query counts accompany headline percentages.

Other endpoints. Stale Citation Rate is the fraction of scorable reader outputs on affected queries that cite any member of Γ_e. Recall@100 is the fraction of queries retrieving at least one admissible

gold support among the first 100 results. Action endpoints measure forbidden-source citation for ATTRBAN, sanctioned-successor Promotion@20 for REPLACE, within-window suppression and post-expiry Recall@20 for TEMPWITH. Answer disclosure is the fraction of fixed-reader responses containing an adjudicated forbidden proposition; lawful-answer support uses matched unaffected queries and counts only supported, non-abstaining answers. Update GPU-hours sum active device time across training, encoding, and reinsertion. Handling time is event arrival to an atomically published serving snapshot; serving p95 is request latency after that snapshot.

Comparison protocol. All efficient methods begin from the same frozen Contriever checkpoint and FAISS snapshot, receive identical event/query splits, and tune only on development windows. Five seeds (42, 137, 256, 512, 1024) are paired by window. The shared development stopping rule requires development Exposure@20 to change by at most 0.3 points over three checkpoints; test Exposure never controls stopping. Because packages touch different amounts of model and index state, retrieval quality is reported together with active GPU-hours and elapsed handling time. We report mean $\pm$ standard deviation across seeds and paired event/query bootstrap 95% confidence intervals for prespecified Wikipedia hard-slice differences.

4 ESTIMATING AND EDITING THE CLOSURE WITH PRAU

PRAU separates three operations that exact deletion conflates: resolving the event's reachable neighborhood, deciding admissibility under the requested operator, and editing only the document representations that must change. The benchmark closure Γ_e remains hidden; the updater acts on its event-visible estimate $\widehat{\Gamma}_e$.

Compatible local representations. PRAU freezes the base query and document towers, $f_{\theta_q^0}$ and $g_{\theta_d^0}$, and learns an event-local document residual a_ψ only for the resolved neighborhood:

$$\begin{aligned} u_q &= f_{\theta_q^0}(q), \\ v_e(d,t) &= g_{\theta_d^0}(d) + \mathbb{1}[d \in \widehat{\Gamma}_e] a_\psi(d,e,t), \\ s_{\mathrm{rel}}(q,d,e,t) &= u_q^\top v_e(d,t), \qquad s_{\mathrm{adm}}(d,e,t) = \sigma(h_\phi(r_e, \mathbf{p}_{d,e,t})), \\ s(q,d,e,t) &= \max(0, s_{\mathrm{rel}}(q,d,e,t)) s_{\mathrm{adm}}(d,e,t). \end{aligned} \tag{4}$$

The two-layer, 256-hidden-unit admissibility head receives an action embedding together with source-state, lineage-edge, graph-hop, duplicate, quote-overlap, cosine, event-age, expiry, and successor indicators. A frozen Contriever/FAISS first stage produces a shared development-selected candidate depth M for every method; action-specific hard masks are applied first, Eq. 4 reranks the survivors, and the first k are returned. Clipping retained relevance scores at zero makes decreasing admissibility unable to raise a candidate score; equal admissibility preserves the ordering of positive relevance scores. Freezing both base towers leaves untouched and locally edited vectors in the same representation space.

Resolved neighborhood. The provenance graph $\mathcal{G} = (\mathcal{C}_t, \mathcal{E})$ connects chunks by four edge types: URL lineage/canonical source, MinHash duplicate membership (≥ 0.8), quote overlap (> 30 tokens), and embedding cosine (> 0.92). Let $D(S_e)$ be the exact chunks resolved from the event sources. PRAU constructs the event-time closure estimate $\widehat{\Gamma}_e = D(S_e) \cup \mathcal{N}_2(D(S_e))$, applies the immediate action policy to exact chunks and duplicate clusters, and uses the remainder for admissibility scoring, replacement promotion, local fine-tuning, and reinsertion. Hard negatives include mirrors, quote fragments, lightly rewritten copies, and embedding neighbors. The separation between independently adjudicated Γ_e and updater-visible $\widehat{\Gamma}_e$ makes closure coverage measurable without sharing evaluation labels with the method.

Objective and same-input control. Each episode contains surviving positives D^+_{surv}, revoked negatives D^-_{rev}, and resolved neighbors D_{neigh}:

$$\mathcal{L} = \mathcal{L}_{\mathrm{recall}} + \lambda_1 \mathcal{L}_{\mathrm{adm}} + \lambda_2 \mathcal{L}_{\mathrm{exp}} + \lambda_3 \mathcal{L}_{\mathrm{coll}}, \tag{5}$$

Here $\mathcal{L}_{\mathrm{recall}}$ is an in-batch contrastive loss over surviving evidence, $\mathcal{L}_{\mathrm{adm}}$ is binary cross-entropy on updater-visible action labels, $\mathcal{L}_{\mathrm{exp}}$ enforces a 0.15 score margin, and $\mathcal{L}_{\mathrm{coll}}$ caps legitimate-neighbor

degradation at 0.05. Appendix C defines all four normalized terms and the episode supervision; evaluation labels z_e never enter these losses. We set $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 1.0, 0.3)$ on development windows. Graph Filter-Only receives the identical resolved graph and masks every flagged serving candidate but learns neither the residual nor the admissibility head, providing a package-level contrast between graph-only masking and the learned update.

Work decomposition. A full update touches all N chunks. PRAU instead performs graph resolution over event-local edges, $T_{\text{update}} = 500$ optimization steps on sampled neighborhood episodes, re-encodes $|\mathcal{N}_2(D(S_e))|$ chunks with the learned residual, and reinserts those vectors in one atomic snapshot swap. The observed neighborhood and reinsertion volume account for the measured update cost. Algorithm 1 and the exact losses appear in Appendix C.

5 EXPERIMENTS

Evaluation questions. The study tests three linked decisions: (Q1) when does an event-local update overtake broader reconstruction in exposure, recall, and compute; (Q2) which parts of closure estimation and admissibility editing account for that trade-off; and (Q3) whether the updated snapshot arrives early enough to change action-specific and reader-level outcomes.

5.1 STREAMS, SPLITS, AND COMPARED SYSTEMS

Wikipedia stream. English Wikipedia dumps from January 2023 through December 2024 yield 10.2M 100-word chunks. Revision deletions, rollbacks, copyright removals, and source replacements define candidate events; the fixed action mixture is 35% FULLDEL, 20% REDACT, 10% ATTRBAN, 25% REPLACE, and 10% TEMPWITH. Standard windows contain 200–500 events and 2K–8K affected chunks. The evaluation includes 2,847 Natural Questions items whose gold evidence intersects an affected source (Kwiatkowski et al., 2019). Stress slices are mirror-heavy (≥ 10 near-duplicates per source; 847 events), bursty (500 sources and approximately 12K chunks in one arrival window), and authority-skewed (312 events on the top 1% of pages by incoming-link degree).

Technical-document stream. We align the October 2024 and October 2025 FreshStack/DriftBench snapshots over ten LangChain-ecosystem repositories (Thakur et al., 2025; Kuissi et al., 2026). Commit/file history resolves removed, deprecated, migrated, and replaced documentation into 18,420 revoked chunks and 2,731 duplicate/migration clusters across 412 events. All 203 public queries are retained: 126 intersect changed support and 77 are lawful controls. Native stress slices contain 46 migration, 43 restructuring, and 37 authority-skewed affected queries.

Methods and implementation. The comparison includes Delete-Only, Continual FT, Hybrid Filter, Retrieval Unlearning (Neel et al., 2021), SISA-RETRIEVAL with 20 fixed shards (Bourtoule et al., 2021), Partial Rebuild, CoCoL (Hou et al., 2026b), CuRD (Hou et al., 2026a), Graph Filter-Only, PRAU, and Full Rebuild. Appendix D specifies the event information, learned state, graph access, and index work of each system package. Headline runs use Contriever (Izacard et al., 2022), AdamW at 2×10^{-5}, batch size 128, CUDA 12.4, and PyTorch 2.3.1 on 8×A100 80GB GPUs. The fixed Llama-3.1-70B reader uses identical prompts and decoding across PRE/POST/FILT states (Grattafiori et al., 2024).

Temporal split and information boundary. Within each stream, canonical URL lineage, duplicate membership, and migration links are collapsed into provenance components before chronological train/development/test assignment; an entire component belongs to one split. Updaters at time τ can use only event-visible corpus state and graph edges timestamped no later than τ. Test queries, candidate-pool judgments, closure labels, and endpoint outcomes are excluded from episode sampling, hard-negative mining, stopping, and selection of the 0.92/two-hop operating point. Standard and stress slices are views of the held-out stream, while typed-action and answer subsets are endpoint-specific views of the same affected-query pool rather than independent replications.

Table 1: The Partial Rebuild–PRAU reversal. Entries are Exposure@20 (%, lower is better); bold marks the better of these two update strategies. SISA-RETRIEVAL gives a shard-retraining reference. Detailed uncertainty, compute, and additional comparators appear in Tables 2, 7, and Appendix A.

Stream	Setting	SISA	Partial	PRAU
Wikipedia	Standard	11.3	**8.7**	9.7
	Mirror-heavy	22.4	16.9	**14.2**
	Bursty	27.6	19.4	**12.8**
	Authority	19.8	15.8	**11.6**
Technical docs	Standard	12.1	**8.8**	9.6
	Migration	23.7	17.2	**15.6**
	Restructure	29.4	18.8	**14.8**
	Authority	20.9	15.3	**13.5**

Table 2: Wikipedia high-churn Exposure@20/Recall@100 (%, lower/higher is better), mean ± std over five paired seeds. Underlining marks the strongest non-rebuild pair; update compute appears in Table 6.

Method	Mirror-heavy	Bursty	Authority
Delete-Only	56.3±1.8/67.8±0.7	62.8±2.1/64.2±0.8	51.4±1.6/68.9±0.6
Continual FT	34.7 ± 1.5/77.0 ± 0.6	39.9 ± 1.8/75.1 ± 0.8	31.6 ± 1.3/77.8 ± 0.6
Hybrid	25.8 ± 1.2/79.4 ± 0.5	31.7 ± 1.5/76.9 ± 0.7	23.6 ± 1.0/79.8 ± 0.5
Retr. Unlearn.	28.7±1.4/74.1±0.6	34.1±1.7/71.8±0.8	26.3±1.2/75.2±0.5
SISA-RETRIEVAL	22.4±1.1/76.3±0.5	27.6±1.3/73.4±0.7	19.8±0.9/77.1±0.5
Partial	16.9 ± 0.9/81.1 ± 0.5	19.4 ± 1.1/80.2 ± 0.6	15.8 ± 0.8/81.4 ± 0.4
CoCoL	20.6 ± 1.0/80.0 ± 0.5	23.9 ± 1.2/78.7 ± 0.6	18.4 ± 0.8/80.5 ± 0.5
CuRD	18.2 ± 0.9/80.8 ± 0.5	21.5 ± 1.1/79.6 ± 0.6	16.7 ± 0.8/81.0 ± 0.4
Graph Filter	17.6 ± 0.9/80.6 ± 0.5	20.8 ± 1.1/79.5 ± 0.6	16.2 ± 0.8/80.9 ± 0.5
PRAU	14.2±0.8/81.7±0.5	12.8±0.7/82.3±0.4	11.6±0.6/82.1±0.4
Full Rebuild	3.4±0.4/83.8±0.3	2.9±0.3/84.1±0.3	4.1±0.5/83.5±0.4

5.2 ORDINARY CHURN FAVORS BROADER RECONSTRUCTION

On the standard Wikipedia control over all 2,847 affected queries, Full Rebuild reaches 2.1±0.3% Exposure@20 and 84.6±0.3 Recall@100. Efficient methods cluster much closer: PRAU reaches 9.7±0.5 exposure, 5.8±0.4 stale citation, and 83.4±0.4 recall at 5.8 GPU-hours. Partial Rebuild buys a 1.0-point exposure reduction (8.7 versus 9.7) at 18.7 GPU-hours—3.2× the update cost—and raises stale citation from 5.8 to 9.4. Table 5 quantifies the compressed standard-churn ranking before the stress slices.

5.3 STRESS REVERSES THE PARTIAL-REBUILD ORDERING

Table 2 reports Exposure@20/Recall@100 for the full comparator set. Mirror-heavy, bursty, and authority-skewed slices contain 1,936, 2,214, and 1,502 affected queries, corresponding to 38,720, 44,280, and 30,040 audited top-20 results per seed.

The strategy ordering in the ordinary control is not stable under closure stress. Partial Rebuild's 1.0-point advantage becomes a 2.7-point deficit on mirror-heavy events, 6.6 points on bursts, and 4.2 points on authority-skewed events; it also uses 20.8 ± 0.8 GPU-hours, versus 6.4 ± 0.3 for PRAU. Among methods averaging at most 10 GPU-hours per hard-slice cycle, PRAU therefore defines the strongest closure-quality point. Relative to SISA-RETRIEVAL, it reduces exposure by 8.2 points on mirror-heavy events, 14.8 on bursts, and 8.2 on authority-skewed events; paired 95% confidence intervals are $[7.5, 8.9]$, $[13.9, 15.7]$, and $[7.6, 8.8]$. At the median seed, this is 275/1,936, 283/2,214, and 174/1,502 exposed queries, compared with 434, 611, and 297 for SISA-RETRIEVAL. CoCoL and CuRD edit specified query–document associations, whereas PRAU combines event-local provenance resolution, operator-conditioned gating, and document residuals. Graph Filter-Only receives the same graph but remains 3.4–8.0 points higher in exposure. Paired intervals against CuRD, Graph

Table 3: Endpoint-matched action evaluation (%, mean ± std, five seeds). Attr citation is lower-better; the other endpoints are higher-better.

Method	Attr citation	Replace P@20	Temp suppress	Reinstate R@20
SISA-RETRIEVAL	12.9 ± 0.7	78.1 ± 0.7	84.8 ± 0.8	86.7 ± 0.7
CuRD	10.8 ± 0.7	82.4 ± 0.6	88.0 ± 0.7	88.6 ± 0.6
PRAU	6.7 ± 0.5	86.9 ± 0.6	92.4 ± 0.5	90.8 ± 0.6
Full Rebuild	3.5 ± 0.3	89.4 ± 0.5	94.1 ± 0.4	94.0 ± 0.4

Filter-Only, and Partial Rebuild remain positive on every slice, and all corresponding prespecified tests survive Holm correction (Appendix G).

5.4 COMPONENT SENSITIVITY AND RESIDUAL REACH

The standard-churn component study separates exposure sensitivity from collateral utility within the integrated updater. Removing the exposure margin raises Exposure@20 from 9.7 to 16.8, while removing the provenance graph raises it to 14.3. Removing collateral regularization changes exposure only from 9.7 to 10.2 but lowers Recall@100 from 83.4 to 80.7. The operating sweep places 0.92 cosine and two hops near the measured exposure–recall knee: nearby settings lower exposure by 0.6–1.8 points while giving up 0.9–2.4 recall points (Figure 1). These ablations characterize component sensitivity in the standard regime; Table 2 evaluates the complete updater under stress.

Residual exposure is concentrated in transformations beyond the event-local graph. In a blinded, doubly annotated sample of 337 exposed query results, 105/337 are deep paraphrases (31.2%), 79/337 are cross-language transformations (23.4%), 63/337 are cascaded quotations (18.7%), 50/337 are late mirrors (14.8%), and 40/337 are embedding blind spots (11.9%), with Cohen's $\kappa = 0.82$. Held-out paraphrase, translation, delayed-insertion, and long-chain variants preserve this ordering (Appendix F).

5.5 TRANSFER TO EVOLVING TECHNICAL DOCUMENTATION

The second stream preserves the same method initialization, five-seed protocol, handling-window rule, and metrics. Table 1 foregrounds the replicated Partial Rebuild–PRAU reversal, and Appendix Table 7 reports the core comparator quality–cost points. Partial Rebuild leads by 0.8 exposure point on standard updates, then trails by 1.6 on migration, 4.0 on restructuring, and 1.8 on authority-skewed deprecations. Against SISA-RETRIEVAL, the corresponding stress-slice advantages are 8.1, 14.6, and 7.4 points. Mean exposed-query incidences per seed are 7.2/46, 6.4/43, and 5.0/37 for PRAU, versus 10.9, 12.6, and 7.7 for SISA-RETRIEVAL. The same ordering under different content, query provenance, document lifecycles, and corpus scale supports transfer of the pre-rebuild phenomenon.

5.6 ACTION AND ANSWER ENDPOINTS

Table 3 evaluates the three action families that require endpoints beyond exposure: attribution suppression, sanctioned replacement, and temporary withdrawal with reinstatement. The AttrBan, Replace, and TempWith subsets contain 83 events/291 queries, 176 events/704 queries, and 71 events/279 queries, sampled from the fixed action mixture over the 2,847-query pool. PRAU improves every reported action endpoint over CuRD and SISA-RETRIEVAL: citation exposure falls to $6.7 \pm 0.5\%$, sanctioned replacements reach $86.9 \pm 0.6\%$ Promotion@20, temporary suppression reaches $92.4 \pm 0.5\%$, and post-expiry recall reaches $90.8 \pm 0.6\%$. At the median seed these endpoints correspond to 19/291, 612/704, 258/279, and 253/279 query incidences.

The FBD-style answer audit uses a stratified draw of 512 scorable queries from the three hard slices and 512 matched lawful controls with fixed prompts and decoding. A shared PRE state confirms answerability at $82.7 \pm 0.6\%$ disclosure; POST scores the raw updated retriever, and FILT applies its serving policy. Table 4 reports the paired POST→FILT change alongside FILT citation exposure, abstention, and lawful-answer support. PRAU reduces disclosure to $4.9 \pm 0.4\%$, versus $9.7 \pm 0.5\%$ for SISA-RETRIEVAL and $7.6 \pm 0.5\%$ for CuRD, while preserving $81.6 \pm 0.5\%$ lawful support. At the median seed, this is 25/512 disclosures and 418/512 lawfully supported controls.

Table 4: Hard-slice answer audit (%, mean ± std over five seeds). Disclosure is POST/FILT; citation exposure, abstention, and lawful support are FILT-state endpoints. Lower/lower/lower/higher is better.

Method	Disclosure P/F	Cite exposure	Abstain	Lawful support
Delete-Only	$24.6 \pm 1.0/21.8 \pm 0.9$	28.6 ± 1.0	2.9 ± 0.3	84.3 ± 0.5
FBD-FILT	$21.9 \pm 0.9/13.8 \pm 0.7$	10.4 ± 0.6	11.7 ± 0.6	79.9 ± 0.6
SISA-RETRIEVAL	$11.3 \pm 0.6/9.7 \pm 0.5$	8.4 ± 0.5	6.8 ± 0.4	81.1 ± 0.5
CuRD	$8.8 \pm 0.5/7.6 \pm 0.5$	6.9 ± 0.4	6.1 ± 0.4	81.4 ± 0.5
PRAU	$5.6 \pm 0.4/4.9 \pm 0.4$	4.6 ± 0.4	5.4 ± 0.3	81.6 ± 0.5
Full Rebuild	$3.2 \pm 0.3/2.8 \pm 0.3$	3.1 ± 0.3	4.8 ± 0.3	82.4 ± 0.4

5.7 RESPONSE WITHIN THE HANDLING WINDOW

Stage timing over paired burst events gives PRAU an end-to-end p50/p95 of $41.0\,min/46.7\,min$; Full Rebuild requires $6.3\,h/6.8\,h$ and 48.2 GPU-hours. Exact masks are published first, local optimization and reinsertion run against a shadow snapshot, and completion means that the updated snapshot has been atomically swapped into service. At one hour, completion is $98.2 \pm 0.7\%$ for PRAU, $99.8 \pm 0.1\%$ for Graph Filter-Only, $89.6 \pm 2.4\%$ for CuRD, $82.1 \pm 2.8\%$ for SISA-RETRIEVAL, and $71.4 \pm 2.6\%$ for Partial Rebuild. Completion incidence is the all-event timing summary; the accompanying 12.8±0.7 Exposure@20 and 82.3±0.4 Recall@100 values describe completed PRAU event/query pairs. Across 15/30/60/120-minute budgets, PRAU completion rises from $1.4 \pm 0.4\%$ to $18.7 \pm 1.8\%$, $98.2 \pm 0.7\%$, and $99.6 \pm 0.2\%$. Appendix H reports the stage decomposition and conditional boundary metrics.

6 DISCUSSION: WHY STRESS CHANGES THE EFFICIENT UPDATE

The decisive comparison is between reconstruction breadth and closure-aware locality. Partial Rebuild has the best ordinary-churn exposure among efficient methods, but PRAU overtakes it on every stress slice while using roughly one third of its Wikipedia hard-slice compute. The SISA-RETRIEVAL–PRAU gap simultaneously grows from 1.6 points to 8.2/14.8/8.2 on Wikipedia and from 2.5 points to 8.1/14.6/7.4 on technical documents. This pattern is consistent with the geometry of the slices: mirror-heavy events present more derivative routes, bursts create wider simultaneous neighborhoods, and high-authority sources connect stale evidence to more queries. The same-graph control and standard-regime ablations sharpen that account: graph-only masking remains above the full updater, removing the exposure margin raises exposure, and removing the collateral term lowers recall.

Interpretation boundary. The evidence establishes a strategy reversal for authenticated changes in two versioned text streams, candidate-bounded closures, Contriever with an E5 replication, and the reported reader. The stress slices are descriptive operating regimes rather than a multivariable causal decomposition, while Graph Filter-Only is a same-input package control rather than a one-factor ablation of the learned update. Full Rebuild anchors the resolved-corpus endpoint, and completion incidence remains distinct from conditional retrieval quality at each deadline. These boundaries leave the main decision intact: before rebuild, derivative-rich events favor closure-aware local editing, and the measured residual is concentrated in deep paraphrases, cross-language transformations, long quotation chains, and late mirrors.

7 CONCLUSION

Revocation before rebuild is a distinct retrieval problem: the serving state must change within one hour even when complete reconstruction takes 6.8 hours. REVOCEDGE-10M makes this interval measurable through operator-specific, candidate-bounded semantic closure. Partial Rebuild leads PRAU under ordinary churn, yet PRAU reverses that ordering on all six stress slices across Wikipedia and evolving technical documents while requiring substantially less update compute. Its 46.7-minute p95, 4.9% hard-slice answer disclosure, and 81.6% lawful support show that provenance-resolved local editing can turn closure reach into an operational serving update. Dense-retrieval maintenance

should therefore evaluate what answer-bearing evidence remains reachable at the deadline, not only which listed records were removed.

AI USE STATEMENT

Generative AI tools were used to assist manuscript preparation, including language revision, organization, reference formatting, and submission-format checks.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. L2R: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 183–192, 2023. doi: 10.1145/3583780.3614947.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Laura Dietz, Jiangui Chen, Michael Schuhmacher, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Continual learning for generative retrieval over dynamic corpora. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 306–315, 2023. doi: 10.1145/3583780.3614821.

Antonio Ginart, Melody Y Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cb79f8fa58b91d3af6c9c991f63962d3-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842, 2020. URL https://proceedings.mlr.press/v119/guo20c.html.

Jingrui Hou, Axel Finke, and Georgina Cosma. Neural corrective machine unranking. *Information Sciences*, 745:123366, 2026a. doi: 10.1016/j.ins.2026.123366.

Jingrui Hou, Axel Finke, and Georgina Cosma. Neural machine unranking. *IEEE Transactions on Neural Networks and Learning Systems*, 37(6):2782–2793, 2026b. doi: 10.1109/TNNLS.2025.3639808.

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

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

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

Waqas Khan, Tabinda Sarwar, Jingyue Cong, Xun Yi, and Estrid He. Graph-guided selective unlearning for language models: Controlling support routes beyond forget seeds. *arXiv preprint arXiv:2608.26743*, 2026. URL https://arxiv.org/abs/2608.26743.

Nathan Kuissi, Suraj Subrahmanyan, Nandan Thakur, and Jimmy Lin. Still fresh? evaluating temporal drift in retrieval benchmarks. *arXiv preprint arXiv:2603.04532*, 2026. URL https://arxiv.org/abs/2603.04532.

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. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Allan Mukkuzhi, Noella Uwayisenga, Yeabsira Bizualem, Jacob Kammerzell, and Chandranil Chakraborttii. Ghost echoes: Semantic erasure failure in retrieval-backed applications. *arXiv preprint arXiv:2608.20352*, 2026. URL https://arxiv.org/abs/2608.20352.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 931–962, 2021. URL https://proceedings.mlr.press/v132/neel21a.html.

Arya Raeesi and Hanna Roed. Auditing forgetting in limited memory language models. *arXiv preprint arXiv:2607.00605*, 2026. URL https://arxiv.org/abs/2607.00605.

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

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18075–18086, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9627c45df543c816a3ddf2d8ea686a99-Abstract.html.

Leila Tavakoli and Mark Sanderson. Deletion isn't enough: Auditing RAG for selective forgetting. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2721–2733, 2026. doi: 10.1145/3805712.3808545.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. FreshStack: Building realistic benchmarks for evaluating retrieval on technical documents. *arXiv preprint arXiv:2504.13128*, 2025. URL https://arxiv.org/abs/2504.13128.

Bin Wang, Fan Wang, Pingping Wang, Jinyu Cong, Yang Yu, Yilong Yin, Zhongyi Han, and Benzheng Wei. Agentic unlearning: When LLM agent meets machine unlearning. *arXiv preprint arXiv:2602.17692*, 2026. URL https://arxiv.org/abs/2602.17692.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. URL https://arxiv.org/abs/2212.03533.

Table 5: Standard-churn Wikipedia results. Exposure@20 and Stale Citation Rate are lower-better; Recall@100 is higher-better. Mean ± std over five seeds.

Method	Exposure	Stale cite	Recall	GPU-h	p95 ms
Delete-Only	38.4±1.1	24.7±0.9	71.2±0.4	0.3	18.2
Continual FT	18.3±0.8	14.1±0.6	79.8±0.6	6.4	18.1
Hybrid Filter	15.6±0.7	11.8±0.5	80.5±0.4	4.1	22.7
Retrieval Unlearning	12.4±0.9	10.2±0.8	77.3±0.7	7.8	18.3
SISA-RETRIEVAL	11.3±0.7	8.4±0.5	79.6±0.5	9.2	18.7
Partial Rebuild	8.7±0.6	9.4±0.7	82.1±0.5	18.7	18.4
PRAU	9.7±0.5	5.8±0.4	83.4±0.4	5.8	19.1
Full Rebuild	2.1±0.3	3.2±0.4	84.6±0.3	48.2	17.9

Table 6: Mean GPU-hours per hard-slice cycle over five seeds. Full Rebuild retains the matched 48.2 GPU-hour total.

Method	GPU-h	Method	GPU-h
Delete-Only	0.4 ± 0.1	Continual FT	6.7 ± 0.3
Hybrid Filter	4.3 ± 0.2	Retrieval Unlearning	8.1 ± 0.4
SISA-RETRIEVAL	9.6 ± 0.4	Partial Rebuild	20.8 ± 0.8
CoCoL	7.7 ± 0.3	CuRD	8.5 ± 0.4
Graph Filter-Only	1.4 ± 0.1	PRAU	6.4 ± 0.3
Full Rebuild	48.2		

A STANDARD-CHURN CONTROL

B TECHNICAL-DOCUMENT RESULTS

Table 7 expands the compact transition summary in Table 1. The standard and three native stress slices use the same five paired seeds and report update compute separately from retrieval quality.

Table 7: FreshStack/DriftBench results. Exposure@20 (Exp), Recall@100 (R), and GPU-hours (Cost) are mean ± std over five seeds; lower/higher/lower is better.

Setting	Method	Exp	R	Cost
Standard	SISA-RETRIEVAL	12.1 ± 0.8	83.0 ± 0.5	5.0 ± 0.2
	Partial Rebuild	8.8 ± 0.6	85.1 ± 0.4	8.9 ± 0.4
	CuRD	10.7 ± 0.6	84.2 ± 0.5	5.5 ± 0.3
	PRAU	9.6 ± 0.5	85.8 ± 0.4	4.1 ± 0.2
	Full Rebuild	2.4 ± 0.3	87.1 ± 0.3	19.4 ± 0.7
Migration	SISA-RETRIEVAL	23.7 ± 1.2	80.2 ± 0.6	5.6 ± 0.3
	Partial Rebuild	17.2 ± 0.9	84.4 ± 0.5	9.8 ± 0.4
	CuRD	19.1 ± 1.0	82.6 ± 0.5	6.1 ± 0.3
	PRAU	15.6 ± 0.8	84.7 ± 0.5	4.6 ± 0.2
	Full Rebuild	3.9 ± 0.4	86.4 ± 0.4	19.4 ± 0.7
Restructure	SISA-RETRIEVAL	29.4 ± 1.4	78.5 ± 0.7	6.1 ± 0.3
	Partial Rebuild	18.8 ± 1.0	83.5 ± 0.6	10.7 ± 0.5
	CuRD	22.6 ± 1.1	81.1 ± 0.6	6.5 ± 0.3
	PRAU	14.8 ± 0.8	84.2 ± 0.5	4.9 ± 0.2
	Full Rebuild	3.2 ± 0.3	86.8 ± 0.3	19.4 ± 0.7
Authority	SISA-RETRIEVAL	20.9 ± 1.0	82.0 ± 0.5	5.5 ± 0.3
	Partial Rebuild	15.3 ± 0.8	85.0 ± 0.4	9.5 ± 0.4
	CuRD	17.4 ± 0.8	83.5 ± 0.5	6.0 ± 0.3
	PRAU	13.5 ± 0.7	85.1 ± 0.4	4.5 ± 0.2
	Full Rebuild	4.4 ± 0.5	86.2 ± 0.4	19.4 ± 0.7

The exposure gap is small in the technical standard slice and largest in restructuring, matching the ordering in the Wikipedia standard and burst settings without equating the two corpora's absolute scales.

C Complete Update Procedure

An update episode pairs an event-visible training query with surviving positives D^{+}_{surv}, exact or graph-resolved revoked negatives D^{-}_{rev}, legitimate resolved neighbors D_{neigh}, and in-batch background negatives. Surviving positives are pre-event supports whose action state remains admissible; revoked and admissible targets are derived only from the authenticated event and updater-visible provenance graph. The evaluation judgment $z_e(d,t)$ and all test queries are excluded. For episode batch $\mathcal{B}$, candidate set $C(q,e)$, temperature η, and updater target $\widehat{y}_e(d,t) \in \{0,1\}$, the normalized terms in Eq. 5 are

$$\mathcal{L}_{\mathrm{recall}} = -\frac{1}{|\mathcal{B}|} \sum_{(q,e)\in\mathcal{B}} \log \frac{\sum_{d^+\in D^{+}_{\mathrm{surv}}} \exp(s(q,d^+,e,t)/\eta)}{\sum_{d\in C(q,e)} \exp(s(q,d,e,t)/\eta)}, \tag{6}$$

$$\mathcal{L}_{\mathrm{adm}} = -\frac{1}{|\mathcal{B}|} \sum_{(q,e)\in\mathcal{B}} \frac{1}{|C(q,e)|} \sum_{d\in C(q,e)} \left[\widehat{y}_e(d,t)\log s_{\mathrm{adm}} + (1-\widehat{y}_e(d,t))\log(1-s_{\mathrm{adm}})\right], \tag{7}$$

$$\mathcal{L}_{\mathrm{exp}} = \frac{1}{|\mathcal{B}|} \sum_{(q,e)\in\mathcal{B}} \frac{1}{|D^{-}_{\mathrm{rev}}|} \sum_{d^-\in D^{-}_{\mathrm{rev}}} \max\left(0,\ s(q,d^-,e,t) - \max_{d^+\in D^{+}_{\mathrm{surv}}} s(q,d^+,e,t) + \gamma\right), \tag{8}$$

$$\mathcal{L}_{\mathrm{coll}} = \frac{1}{|\mathcal{B}|} \sum_{(q,e)\in\mathcal{B}} \frac{1}{|D_{\mathrm{neigh}} \setminus D^{-}_{\mathrm{rev}}|} \sum_{d\in D_{\mathrm{neigh}}\setminus D^{-}_{\mathrm{rev}}} \max(0,\ s_{\mathrm{pre}}(q,d) - s(q,d,e,t) - \delta). \tag{9}$$

Here s_{adm} abbreviates $s_{\mathrm{adm}}(d,e,t)$ within the inner sum, $\gamma = 0.15$, and $\delta = 0.05$. Empty per-episode sets contribute zero. Episodes are sampled from training windows, the operating point and stopping checkpoint are selected on development windows, and held-out event components enter evaluation only.

Algorithm 1 PRAU: Provenance-Resolved Admissibility Update

Require: Event $e = (\tau, S_e, r_e)$, graph $\mathcal{G}$, frozen towers θ^0_q, θ^0_d, residual ψ, head ϕ
1: Apply the immediate action policy to exact chunks and known duplicates
2: Mask removed/withdrawn evidence; redact spans or attribution as specified
3: Resolve $\mathcal{N}_2(D(S_e))$ through URL, duplicate, quote, and cosine edges
4: Score $s_{\mathrm{adm}}(d,e,\tau) \leftarrow \sigma(h_\phi(r_e, \mathbf{p}_{d,e,\tau}))$
5: **if** $r_e =$ REPLACE **then**
6: Promote the sanctioned successor d'
7: **end if**
8: Sample surviving, revoked, and collateral-neighbor episodes
9: Update ψ, ϕ with Eq. 5 for 500 steps
10: Encode $v_e(d,\tau)$ for $\mathcal{N}_2(D(S_e))$ and reinsert changed vectors
11: Atomically publish the masked and locally updated snapshot

D Baseline Adaptations and Information Access

Table 8 makes the comparison unit explicit. Each row is a complete updater package rather than a one-factor intervention. Gradient-based packages use their published objective adapted to the shared Contriever representation, development-only hyperparameter selection, and the stability rule in Section 3; none receives A_e, Γ_e, or test queries.

GPU-hours count active optimization, encoding, and reinsertion; elapsed handling time includes graph resolution and the atomic snapshot publication. Persistent graph construction is shared background state for the two graph-visible packages, so its event-time lookup is counted but initial offline construction is not charged to either update.

Table 8: Baseline adaptation contract. "Event" is the authenticated exact source/chunk action; "graph" is the updater-visible PRAU provenance graph. All rows start from the same checkpoint and index and use the same first-stage candidate depth.

Package	Visible update information	Graph	Learned state	Index work
Delete-Only	Exact event members	No	None	Apply exact action policy
Continual FT	Event forget/retain pairs	No	Method-specific adapter	Re-encode method-selected documents
Hybrid Filter	Event terms and exact IDs	No	None	Sparse filter before dense ranking
Retr. Unlearning	Exact forget/retain pairs	No	Method-specific adapter	Re-encode method-selected documents
SISA-RETRIEVAL	Event and fixed shard map	No	Affected shard models	Retrain and re-encode touched shards
Partial Rebuild	Resolved touched shards	No	None	Re-encode touched shards
CoCoL	Exact forget/retain associations	No	Unranking update	Re-encode updated documents
CuRD	Forget/retain pairs and successor	No	Corrective unranking update	Re-encode updated documents
Graph Filter-Only	Event and identical graph	Yes	None	Mask graph-flagged candidates
PRAU	Event, action, and identical graph	Yes	Residual and admissibility head	Re-encode resolved neighborhood
Full Rebuild	Resolved post-event corpus	Corpus resolution	Full reference encoder/index	Re-encode and re-index corpus

E OPERATING-POINT AND BACKBONE ROBUSTNESS

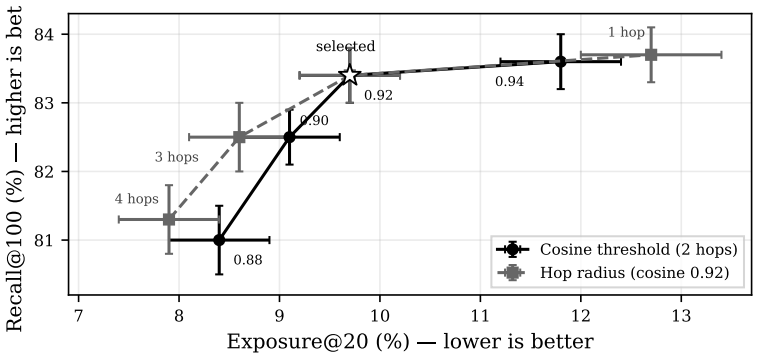

Figure 1: Held-out Wikipedia standard-window operating sweep over five paired seeds (mean ± std; lower exposure and higher recall are better). Threshold and hop-radius paths meet at the development-selected 0.92/two-hop point; marker and line-style differences preserve grayscale readability.

Figure 1 visualizes the development-selected operating point; Table 9 gives the underlying values and the second-backbone table tests whether the stress ordering depends on Contriever.

Table 9: Development-selected graph sweep on standard Wikipedia windows. The held-out 0.92/two-hop point is the backed main result.

Graph point	Exposure@20	Recall@100
0.88 cosine, 2 hops	8.4 ± 0.5	81.0 ± 0.5
0.90 cosine, 2 hops	9.1 ± 0.5	82.5 ± 0.4
0.92 cosine, 1 hop	12.7 ± 0.7	83.7 ± 0.4
0.92 cosine, 2 hops	9.7 ± 0.5	83.4 ± 0.4
0.92 cosine, 3 hops	8.6 ± 0.5	82.5 ± 0.5
0.92 cosine, 4 hops	7.9 ± 0.5	81.3 ± 0.5
0.94 cosine, 2 hops	11.8 ± 0.6	83.6 ± 0.4

At the selected graph point, query-incidence exposure increases with audit depth as expected: Exposure@5/10/20/50 is 4.8 ± 0.4, 7.1 ± 0.4, 9.7 ± 0.5, and 15.8 ± 0.7. Across neighboring graph points, PRAU remains ahead of SISA-RETRIEVAL, CuRD, and Graph Filter-Only on all three hard slices.

Table 10: E5 second-backbone replication. Exposure@20 (Exp) and Recall@100 (R) use the same five-seed Wikipedia protocol and denominators as the Contriever study.

Setting	Method	Exp	R
Standard	SISA-RETRIEVAL	11.7 ± 0.7	80.4 ± 0.5
	CuRD	10.9 ± 0.6	84.0 ± 0.4
	PRAU	9.2 ± 0.5	85.0 ± 0.4
	Full Rebuild	2.0 ± 0.3	86.1 ± 0.3
Mirror	SISA-RETRIEVAL	21.8 ± 1.0	77.1 ± 0.5
	CuRD	17.6 ± 0.9	81.4 ± 0.5
	PRAU	13.7 ± 0.7	82.5 ± 0.4
	Full Rebuild	3.2 ± 0.4	84.7 ± 0.3
Bursty	SISA-RETRIEVAL	26.9 ± 1.2	74.3 ± 0.7
	CuRD	20.9 ± 1.0	80.2 ± 0.6
	PRAU	12.4 ± 0.7	83.0 ± 0.4
	Full Rebuild	2.7 ± 0.3	85.0 ± 0.3
Authority	SISA-RETRIEVAL	19.1 ± 0.9	77.9 ± 0.5
	CuRD	16.1 ± 0.8	81.6 ± 0.4
	PRAU	11.2 ± 0.6	82.8 ± 0.4
	Full Rebuild	3.9 ± 0.4	84.4 ± 0.4

F FAILURE ANNOTATION AND ADAPTIVE VARIANTS

Two annotators independently label a stratified sample of 337 exposed query results using a frozen codebook and no method identity. Disagreements are adjudicated after independent coding. The counts are 105 deep paraphrases, 79 cross-language mirrors, 63 cascaded quotations, 50 late-arriving mirrors, and 40 embedding blind spots, reproducing the 31.2/23.4/18.7/14.8/11.9% decomposition. Overall agreement is $\kappa = 0.82$.

Table 11: Held-out adaptive variants, excluded from graph construction. Exposure@20/Recall@100 (%) is reported over five seeds with raw attack counts.

Variant	Attacks	SISA-RETRIEVAL	PRAU
Semantic paraphrase	400	$32.4 \pm 1.4/77.6 \pm 0.6$	$18.6 \pm 0.9/81.5 \pm 0.5$
Translation/back-translation	400	$33.8 \pm 1.5/76.9 \pm 0.7$	$20.1 \pm 1.0/80.9 \pm 0.5$
Delayed mirror insertion	300	$28.7 \pm 1.2/78.3 \pm 0.5$	$16.3 \pm 0.8/82.0 \pm 0.4$
Four-to-six-hop chain	300	$36.9 \pm 1.6/75.8 \pm 0.7$	$22.8 \pm 1.1/80.7 \pm 0.5$

G ABLATION AND DISTRIBUTIONAL EVIDENCE

Table 12: PRAU component ablation on standard Wikipedia churn. Mean $\pm$ std over five seeds.

Variant	Exposure@20	Recall@100	Stale citation
PRAU	9.7 ± 0.5	83.4 ± 0.4	5.8 ± 0.4
− Exposure margin	16.8 ± 0.7	83.1 ± 0.4	10.4 ± 0.6
− Collateral regularization	10.2 ± 0.5	80.7 ± 0.5	6.1 ± 0.4
− Provenance graph	14.3 ± 0.6	83.0 ± 0.4	8.9 ± 0.5
− Typed action	12.9 ± 0.6	82.8 ± 0.5	8.2 ± 0.5
− Hard-negative mining	13.1 ± 0.7	83.2 ± 0.4	9.1 ± 0.6

Across paired hard windows, PRAU Exposure@20 has median 11.9% and IQR $[9.8,\ 14.6]\%$; its p90 and p95 are 17.8% and 20.6%. The fraction of windows above 20% exposure is $5.7 \pm 0.8\%$, compared with $31.4 \pm 2.1\%$ for SISA-RETRIEVAL. Paired 95% bootstrap intervals for the PRAU exposure reduction against CuRD are $[3.4,\ 4.6]$ on mirror, $[7.9,\ 9.5]$ on burst, and $[4.5,\ 5.7]$ on authority slices. The corresponding intervals are $[2.8,\ 4.0]$, $[7.2,\ 8.8]$, and $[4.0,\ 5.2]$ against Graph Filter-Only, and $[2.1,\ 3.3]$, $[5.8,\ 7.4]$, and $[3.6,\ 4.8]$ against Partial Rebuild. Events are the resampling unit; all 12 prespecified method–slice tests, including the three SISA-RETRIEVAL comparisons in Section 5.3, remain significant after Holm correction.

H TIMING DECOMPOSITION

Table 13: Elapsed stage time over paired burst events. Stage percentiles are computed independently, while end-to-end latency is measured directly; percentile rows are therefore non-additive. GPU-hours are reported separately in Table 6.

Stage	p50	p95
Hard mask	$4.8\,s$	$7.1\,s$
Graph resolution	$41\,s$	$58\,s$
Admissibility scoring	$64\,s$	$82\,s$
Local optimization	$12.1\,min$	$13.4\,min$
Re-encoding and reinsertion	$27.6\,min$	$31.2\,min$
PRAU end-to-end	$41.0\,min$	$46.7\,min$
Full Rebuild end-to-end	$6.3\,h$	$6.8\,h$

Completion and retrieval quality answer different boundary questions. The completion column uses every burst event and records whether the local snapshot is serving by the budget; Exposure and Recall characterize the completed event/query pairs available at that boundary.

Table 14: PRAU completion and quality at declared elapsed-time budgets over burst events. Completion is event incidence; exposure and recall use completed event/query pairs at the boundary.

Budget	Completed (%)	Exposure@20	Recall@100
15 min	1.4 ± 0.4	31.8 ± 1.5	82.8 ± 0.5
30 min	18.7 ± 1.8	20.6 ± 1.1	82.6 ± 0.4
60 min	98.2 ± 0.7	12.8 ± 0.7	82.3 ± 0.4
120 min	99.6 ± 0.2	12.4 ± 0.7	82.2 ± 0.4

The sharp completion increase between 30 and 60 minutes follows the measured local-optimization and reinsertion stages in Table 13; the separate reporting keeps queue completion distinct from conditional retrieval quality.