RESIDUAL RETRIEVAL MEMORY FOR FORGETTING DECISIONS IN HYBRID RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Deciding whether a hybrid retrieval-augmented generation system has forgotten a fact requires testing the distributed state that can reconstruct it, not only the disappearance of one direct answer. On Natural Questions, eight repairs occupy a narrow 0.8–3.1% direct range while spanning 8.2–16.7% paraphrase and 10.4–21.3% multi-hop leakage, producing different stopping decisions from nearly indistinguishable direct scores. We formalize the missing decision object as *residual retrieval memory* and introduce the Residual Retrieval Access Benchmark (RRA-BENCH), which profiles six reconstruction channels, separates query-only attacks from target-bearing reintroduction, and pairs fact compromise with retained utility and repair cost. A crossed study of 150 method–stage–seed cells finds that direct leakage has Spearman $\rho = 0.23$ [0.15, 0.31] with indirect risk and only 0.61 [0.58, 0.65] AUROC for reintroduction-augmented AnyLeak@32; at a 1% direct threshold, 13 of 18 passing cells remain false-safe. PATHSWEEP, a trace-conditioned allocator over dense, sparse, metadata, reranker, and cache surfaces, provides a constructive test under the same operations and budgets as fixed and validation-selected schedules; it lowers Hard Delete's reintroduction-augmented AnyLeak@32 from 38.6% [36.2,41.1] to 18.7% [16.4,21.2] while retaining 0.793 global MRR. Disjoint prompts, independently authored Qwen-2.5 attacks, 72-hour routine ingestion, and a component-replaced HotpotQA stack preserve the residual-risk separation hidden by direct scores. Residual retrieval memory thus preserves the distinctions needed to stop and rank hybrid-RAG repairs after direct answers disappear.

1 INTRODUCTION

A hybrid-RAG forget request changes a decision, not merely an answer. The system must decide when repair can stop and which repair leaves the least recoverable target information. A direct probe observes only the shortest execution path: source retrieval followed by an answer. Dense neighbors, sparse aliases, metadata joins, reranker features, cached summaries, and later ingestion can support different paths to the same relation. Two repairs can therefore pass the same direct check while inducing sharply different fact-reconstruction risk.

This mismatch follows from the architecture. Hybrid retrieval is a sequence of stateful transformations rather than a lookup in one removable document: dense and sparse indexes expose different candidates, rerankers and metadata reconnect evidence, and caches preserve derived context (Guu et al., 2020; Lewis et al., 2020; Izacard & Grave, 2021; Gao et al., 2023). Paraphrase, evidence composition, dialogue, alias translation, and corpus evolution traverse different subsets of that state. An answer-level score can be accurate for its route yet insufficient for the stopping and ranking decisions applied to the complete pipeline.

Dynamic LLM-unlearning probes recover residual knowledge through aliases and multi-hop chains (Shah et al., 2025); MEMLEAK exposes correlated text and image channels after direct deletion and organizes them with an information-provenance graph (Wang & Zhang, 2026). RAG studies measure behavioral suppression, selective record revocation, retrieved exclusions, and database interventions (Wang et al., 2026; Tavakoli & Sanderson, 2026; Kim et al., 2025; Raeesi & Roed, 2026). What remains unmeasured is the criterion-validity estimand for hybrid retrieval: the probability that a protected relation is still reconstructible after a repair passes its direct check.

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.

Residual retrieval memory names that fact-level event: the protected relation can be reconstructed through live retrieval or derived state after direct suppression. RRA-BENCH estimates it through direct, paraphrase, multi-hop, multi-turn, multilingual-alias, and routine-ingestion channels. Query-only risk is reported by channel; four snapshot-isolated target-bearing interventions extend the fixed union to reintroduction-augmented AnyLeak@32. Nearby and global retention, previously correct answers lost, runtime, and edit counts expose the cost of reducing either endpoint. Repairs share facts, manifests, adjudication, and accounting, and are frozen before disjoint and independently authored evaluations.

Leakage traces then become a systems experiment. PATHSWEEP ranks candidate edits by active-path coverage per estimated retain loss across dense, sparse, metadata, reranker, and cache surfaces. Against identical surface access, it reaches lower paraphrase, multi-hop, multi-turn, and routine-drift leakage than fixed, random, and validation-selected schedules while using 126 edits per fact.

Across the crossed design, direct ASR fails the stopping decision in three ways: direct-to-indirect Spearman correlation is 0.23, fact-compromise AUROC is 0.61, and about 70% of cells passing 1–5% direct thresholds remain false-safe. This criterion result anchors the paper. RRA-BENCH instantiates the comparison across 150 crossed cells, two 500-fact stacks, prompt and attacker turnover, blinded adjudication, and synchronized cost; trace-conditioned allocation then turns the same residual paths into guidance for where repair effort is useful. The resulting evidence changes the admissible repair set once direct answers disappear.

2 RELATED WORK

From model unlearning to residual-access tests. Machine unlearning targets the influence of training records (Cao & Yang, 2015; Ginart et al., 2019; Guo et al., 2020; Bourtoule et al., 2021); LLM evaluations extend that target to extraction, factual behavior, contextual use, and edit locality (Carlini et al., 2021; Eldan & Russinovich, 2023; Jang et al., 2023; Maini et al., 2024; Shi et al., 2025; Peng et al., 2026; Mitchell et al., 2022; Meng et al., 2023). The Unlearning Mirage generates alias and multi-hop probes from model-elicited structure and explains their success through alternative internal computation paths (Shah et al., 2025). Our target is external: a protected relation distributed through live hybrid-retrieval and derived state. Accordingly, the central estimand is whether direct post-repair behavior predicts retrieval-mediated compromise, rather than whether a weight update removes internal knowledge.

Closest memory and RAG audits. MEMLEAK is the closest residual-memory benchmark: it shows below-1% direct leakage alongside recovery from correlated text and retained images, and its information-provenance graph classifies deletion affordances in multimodal agent memory (Wang & Zhang, 2026). RRA-BENCH instead tests a decision rule over text hybrid-RAG state: it estimates whether a direct stopping statistic preserves rankings and predicts fact compromise across dense, sparse, metadata, reranker, cache, and ingestion surfaces. Wang et al. optimize retrieval scores and constraint context for behavioral RAG unlearning (Wang et al., 2026); our same-stack instantiation holds outer facts, budgets, evaluator, and timing fixed while evaluating a broader residual-access profile. Selective-forgetting report cards (Tavakoli & Sanderson, 2026) score post-deletion behavior, retrieved-exclusion correction (Kim et al., 2025) alters answer context, and database-intervention audits (Raeesi & Roed, 2026) study bounded memory stores. The distinguishing object here is the validity of a stopping statistic, supported by crossed proxy tests, query-budgeted unions, matched surface interventions, and component turnover.

Retrieval systems and trace attribution. KILT, BEIR, MTEB, and MIRACL standardize retrieval quality across tasks and representations (Petroni et al., 2021; Thakur et al., 2021; Muennighoff et al., 2023; Zhang et al., 2023), while dense, sparse, late-interaction, and retrieval-enhanced models distribute evidence across distinct representations (Karpukhin et al., 2020; Khattab & Zaharia, 2020; Borgeaud et al., 2022; Izacard et al., 2021; 2023; Asai et al., 2024). Here execution paths connect a reconstructed answer to the pipeline state that supplied it. Leave-one-surface tests assess that attribution, and the No-Graph control tests whether path-conditioned scheduling adds value beyond access to the same edit operations.

3 RESIDUAL RETRIEVAL MEMORY AS A DECISION CRITERION

Protected object and adversary. A request $f = (S_f, A_f, E_f, \mathcal{R}_f)$ contains forbidden spans, normalized aliases, linked entities, and nearby retain constraints. The protected object is the target relation expressed by f across retrieval and derived pipeline state. The query-only adversary has black-box access to the live RAG endpoint and may issue direct, paraphrase, multi-hop, multi-turn, and multilingual-alias queries. A stronger reintroduction adversary may additionally submit one target-bearing item through the ordinary ingestion interface. Target-independent routine ingestion is system evolution, not an adversary action. Neither adversary can modify model weights, repair code, or evaluator labels. Table 19 aligns each evaluation setting with its protected object, varied attacker or intervention, and inferential target; matched controls and utility costs accompany the corresponding result tables.

Risk profile. For method m and fact f, let $L_k(m, f)$ be attack success rate for channel k. Exact success reproduces the canonical answer or a normalized alias. Soft success requires the same target entity and relation; topical overlap, neighboring entities, and broader relations are rejected. A response counts as leaked when either rule succeeds, with exact match taking precedence in diagnostics. We report

$$\mathbf{L}(m, f) = (L_{\text{dir}}, L_{\text{para}}, L_{\text{hop}}, L_{\text{turn}}, L_{\text{xling}}, L_{\text{churn}})$$

Temporal drift and target-bearing ingestion use explicitly separated protocols. L_{churn} measures leakage after 72 hours of target-independent routine ingestion; controlled reintroduction instead supplies one of four target-bearing transformations as an adversarial stress test. For any fixed manifest, AnyLeak@B is the fraction of fact–seed pairs recovered at least once within B endpoint queries. The query-only submanifest defines AnyLeak@28_Q from eight paraphrase, eight multi-hop, six multi-turn, and six multilingual-alias queries. The reported fact-compromise endpoint, AnyLeak@32_R, augments those 28 opportunities with four snapshot-isolated reintroduction queries; we retain the shorter "AnyLeak@32" label in tables and mark it as reintroduction-augmented in captions. The shared order interleaves four complete five-family cycles (20 positions), including all four reintroductions, and then appends four paraphrase–multi-hop pairs and two multi-turn–multilingual pairs. Each reintroduction uses a fresh clone of the post-repair state, ingests one transformed target-bearing item, issues its endpoint query, and is then discarded, so it cannot affect another position. Ingestion actions are reported separately from endpoint-query counts. All methods receive the same seed-specific prompts and order, all positions are scored after a success, dialogue state is reset between probes, and state is retained only within a multi-turn probe. Queries-to-first-leak use the earliest successful position among compromised fact–seed pairs; non-compromised pairs are right-censored at B. Channel ASRs and AnyLeak@28_Q answer the query-only threat model; AnyLeak@32_R answers the stronger reintroduction model. Paired utility loss completes the decision target.

The proxy study crosses six repair initializers (Hard Delete, Delete+Rebuild, Lexical Deny, Dense Masking, Retriever FT, and PATHSWEEP), five cumulative checkpoints (post-initializer direct audit, dense, sparse, metadata, and cache/reranker), and five seeds, yielding 150 cells. Checkpoint one records the named initializer. Checkpoints two through five admit the surface operations in Section 4: for the five non-PATHSWEEP trajectories, candidates on each newly admitted surface are ordered by estimated retain loss and then the seed-fixed identifier; PATHSWEEP reranks all admitted candidates by $s(e)$. Every trajectory uses the same candidate generators, retain guardrail, edit cap, and stopping tests. Stage therefore denotes active editable-surface depth, not repeated measurement of an unchanged endpoint. RAG-UL and No-Graph are endpoint controls and are not included in this predeclared proxy grid. For cell c and fact f, d_{cf} is direct success averaged over that fact's direct prompts, y_{cf} is the binary reintroduction-augmented AnyLeak@32 outcome, and D_c and I_c are the fact-macro direct and equal-family indirect ASR. Spearman and Kendall operate on the 150 (D_c, I_c) pairs. AUROC uses the 75000 crossed (c, f) observations, with d_{cf} as score and y_{cf} as label. The candidate stopping score deliberately retains its operational ten-direct-prompt definition rather than receiving 32 replicated opportunities: AUROC tests whether that score forecasts the stronger budgeted endpoint, while rank association and false-safe rates test the five-channel indirect macro, including routine drift. The bootstrap resamples fact identifiers and retains each sampled fact's complete method–stage–seed vector, so repeated interventions are never treated as independent facts.

For a threshold t, a cell is false-safe when $D_c < t$ while $I_c \geq t$. The 1%, 2%, and 5% values are symmetric-threshold stress tests rather than universal deployment defaults. Given policy-selected channel thresholds $\boldsymbol{\tau}$, an AnyLeak threshold τ_A, and retain-loss budget δ, the decision rule accepts repair m when

$$\mathbb{E}_f[L_k(m, f)] \leq \tau_k \; \forall k, \qquad \text{AnyLeak@B}(m) \leq \tau_A, \qquad \Delta U(m) \leq \delta.$$

The benchmark therefore evaluates three distinct properties of direct ASR: cell ranking through Spearman and Kendall, fact-level discrimination through AUROC, and threshold behavior through false-safe rates.

Trace-conditioned surface map. We instantiate $G = (V, E)$ from retrieval and generation traces. Observed nodes are dense chunks, BM25 postings, reranker candidates and scores, cache entries, and metadata keys. Observed edges record execution dependencies: hit-to-candidate, candidate-to-score, score-to-context, and cache-to-generation-prefix. Inferred edges are restricted to alias and entity links used by metadata guards. A path is active when every observed edge appears in a leaking trace and every inferred link is reachable from the query alias closure. After each probe batch, new nodes and observed edges are merged; inferred links are recomputed only for affected entities. This map attributes an empirical leak to editable pipeline state.

4 TRACE-CONDITIONED SURFACE REPAIR

Surface edits. PATHSWEEP turns the trace map into a constructive test of the audit boundary. Direct deletion removes matching chunks from dense and sparse stores; dense-neighborhood repulsion trains the encoder with $\max(0, \cos(h_{\text{neighbor}}, h_{\text{target}}) + 0.2)$; sparse suppression removes postings that both fall in the target alias closure and exceed the validation-selected high-IDF cutoff; metadata guards block alias-linked queries whose candidates enter the selected semantic neighborhood; and cache invalidation purges summaries, precomputed answers, and reranker features that reference the target. The cumulative checkpoints admit direct, dense, sparse, metadata, and cache/reranker operations in turn. At each checkpoint, all admitted candidates remain in one queue, allowing trace coverage to allocate edits across the available surfaces rather than forcing every operation in the reporting order.

Selection and stopping algorithm. For every leaking trace, the procedure (1) collects active paths, (2) assigns each path to its first editable surface to prevent double counting, (3) enumerates edits from every admitted surface touched by those paths, (4) ranks candidates by path coverage per estimated retain loss, (5) tentatively applies the best candidate and replays affected retain queries, (6) commits a guardrail-valid edit, removes its covered paths, and updates affected graph neighborhoods, and (7) repeats until a full probe batch yields no new exact/soft success, the edit budget is exhausted, or every remaining candidate violates the retain guardrail. The allocation score for edit e is

$$s(e) = \frac{\sum_{p: e \in p} w_{k(p)}/|p|}{\widehat{\Delta\text{MRR}}(e) + 10^{-4}},$$

where w_k is the inverse number of active paths from family k, equalizing family mass. For current state θ and edit e,

$$\widehat{\Delta\text{MRR}}(e) = \max\{0, \text{MRR}(\mathcal{R}_{\text{val}}; \theta) - \text{MRR}(\mathcal{R}_{\text{val}}; \theta \oplus e)\},$$

computed by exact counterfactual replay of validation queries whose top-k results or metadata footprint intersects e; the full validation set is replayed before commit. Ties favor lower estimated loss and then a seed-fixed candidate identifier. Candidate maintenance is $O(P \log C + E_\Delta)$ for P active paths, C queued edits, and E_Δ updated edges; index rebuild costs are accounted separately in runtime. The semantic guard radius is cosine distance 0.18, selected on the repair/retain validation split from {0.10,0.14,0.18,0.22}. Final attack banks never enter radius, cost, or edit selection.

Budgets and parity. Every stage enforces

$$\text{MRR}(\mathcal{R}; i) \geq \text{MRR}(\mathcal{R}; i-1) - 0.02,$$

an absolute, stage-local nearby-query MRR budget. Cumulative utility can move by more than 0.02 after composition. Enumeration may propose additional operations, but at most 160 candidate edits can be applied per fact over 6 repair-probe rounds. Each round draws a seed-fixed repair-only mini-batch covering every profile family with equal family weight and applies the same exact/soft union judge used for scoring; repair probes are disjoint from validation, core, disjoint, and independent manifests. The same evaluator, prompt allocation, wall-clock accounting, and stopping checks apply to all methods. No-Graph Multi-Surface receives identical candidate generators and budgets but commits operations in the fixed global order direct, dense, sparse, metadata, and cache/reranker. This contrast estimates the complete trace-conditioned allocation rule—coverage ranking, retain-loss ranking, and stopping—rather than attributing the gap to graph access alone.

5 Criterion-Validity Experimental Design

Fact and retain-set construction. For each stack, a candidate is eligible when the unmodified pipeline retrieves an aligned supporting passage, answers the canonical question correctly under the union judge, and exposes a single target entity–relation–answer triple with adjudicable aliases. We align the forbidden span to that passage and retain linked entities only when annotators agree on the relation. Normalized duplicate triples, duplicate source documents, and examples whose alternative support directly states the target are collapsed before sampling the 500 forget facts. For HotpotQA, the protected relation is the final-answer relation and both supporting hops are retained for attack construction. Nearby retains exclude the complete target triple and its answer aliases while forming four disjoint slices—graph-distance-one, dense-similarity-only, entity-overlap-only, and relation-overlap-only—whose sizes sum to 2,000; the last slice may share the relation predicate but never the protected subject–object pair. Global retains are sampled from the remaining baseline-answerable pool after removing target answer, entity, relation, and source-document overlap. Repair/retain validation queries and reported retain queries are disjoint. Core, disjoint, and independent attacks share protected facts but use non-overlapping prompt manifests, making prompt and attacker transfer the evaluated generalization axes.

Stacks and splits. Stack A uses Natural Questions (Kwiatkowski et al., 2019), Contriever-MSMARCO, Pyserini BM25, MiniLM-L6-v2 reranking, LRU summary caching, and Llama-3.1-70B generation (Izacard et al., 2021; Lin et al., 2021; Grattafiori et al., 2024). It contains 500 aligned forget facts, 2,000 nearby queries, and 5,000 global queries. Stack B uses HotpotQA (Yang et al., 2018), BGE-M3 dense retrieval (Chen et al., 2024), SPLADE-v3 sparse retrieval, bge-reranker-large, LFU caching with per-entity invalidation, and Mistral-Large-Instruct-2407 generation. It contains 500 aligned forget facts, 2000 nearby queries, and 5000 global queries. Both stacks preserve equivalent forget, attack, routine-drift, reintroduction, and retain definitions. The component-replaced study repeats the predeclared validation procedure on Stack B while freezing the cross-stack edit cap, six-round limit, 0.02 retain guardrail, attack budgets, and stopping rule. This design transfers the decision criterion to a new retrieval geometry while keeping representation-specific choices validation-selected.

Prompt allocation and randomness. The core bank contains 50 prompts per fact: 10 direct, 20 paraphrase, and 20 multi-hop. Auxiliary banks contain 3.6 aliases per fact on average, 12 multi-turn prompts, 8 multilingual-alias prompts, and 4 snapshot-isolated reintroduction instances per fact. Five seeds control template sampling, paraphrase decoding, retriever negative sampling, candidate-edit tie breaking, and ingestion order. Llama-3.1-8B generates in-family paraphrases at temperature 1.2, top-p 0.95, and 96 maximum new tokens.

Comparators. Hard Delete removes exact aligned chunks; Delete+Rebuild reconstructs FAISS after the same deletion; Lexical Deny filters normalized target terms before reranking; and Dense Masking assigns aligned normalized-document IDs a score of $-\infty$ after cosine scoring and before dense/sparse candidate fusion. Retriever FT uses the same pairwise repulsion objective as the dense PathSweep operation, 500 updates at learning rate 10^{-5}, retain negatives, and the common MRR guardrail, but receives no sparse, metadata, reranker, or cache edits. The same-stack RAG-UL implementation follows the published retrieval scoring and constraint-context construction of Wang et al. (2026); the outer facts, seeds, six-round budget, evaluator, and timing protocol are matched. No-Graph uses PathSweep's candidate generators in the fixed order defined above. The factorial

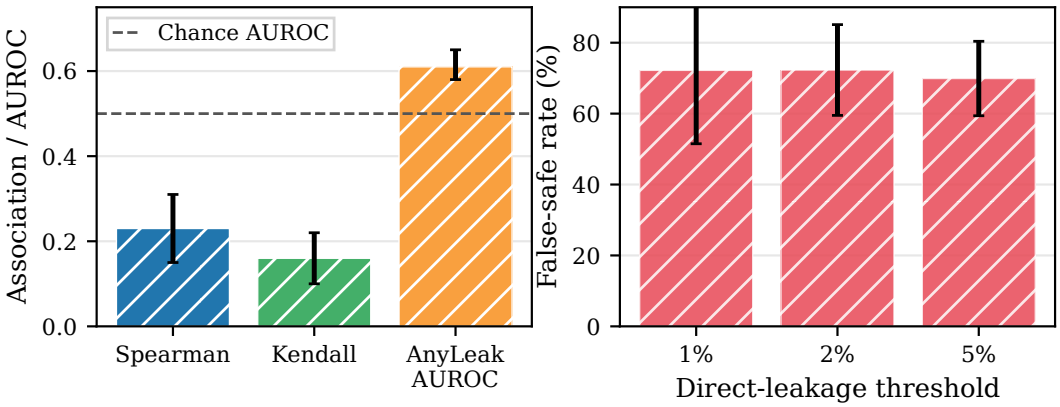


Figure 1: Direct ASR as a stopping surrogate. Rank statistics use 150 cells against the equal-family indirect macro, including routine drift; AUROC uses 75000 crossed cell–fact observations labeled by reintroduction-augmented AnyLeak@32. The right panel reports false-safe cells among 18, 47, and 73 cells passing the 1%, 2%, and 5% direct thresholds. Intervals resample facts while retaining each fact's complete intervention vector.

additionally applies gradient-ascent LM unlearning on target tokens, a context-aware objective (Peng et al., 2026), and sequential retrieval/LM repair. Appendix A.3 gives the operational settings in one place.

Frozen and independent evaluation. After repair, a disjoint bank contributes 7000 direct, 10000 paraphrase, 8000 multi-hop, and 3000 human-authored multi-turn prompts. An independent suite freezes repaired systems and uses Qwen-2.5-32B-Instruct plus method-blind prompt authors for adaptive paraphrase, multi-hop, multi-turn, multilingual-alias, and snapshot-isolated target-bearing reintroduction probes. It covers 500 facts and 6000, 5000, 3500, 4000, and 2000 prompts by family. Adjudicators are blind to method and attack source.

Statistics and runtime. Stochastic attack rates are five-seed means with 95% Student-t intervals. Fixed authored probes, utility, and fact compromise use clustered bootstrap intervals over forget facts or queries grouped by fact. Proxy-validity intervals use the fact-vector resampling defined above; this preserves dependence across methods, checkpoints, and seeds. Runtime is the median of 7 synchronized trials on one forget fact, its 50 core and auxiliary probes, four snapshot-isolated reintroduction instances, and the terminating audit pass, using batch size 16 from the same warm-cache state. Experiments run on one 8×A100-80GB node with CUDA 12.4 and PyTorch 2.3.1.

6 Results

6.1 Weak association makes the direct-leakage floor decision-insufficient

The criterion test precedes the method comparison. Figure 1 asks whether direct ASR preserves either the ranking or the fact-level decisions induced by residual retrieval memory.

Direct ASR weakly orders cell-level indirect risk (Spearman ρ =0.23 [0.15, 0.31]; Kendall τ =0.16 [0.10, 0.22]) and provides limited discrimination of fact compromise (AUROC 0.61 [0.58, 0.65]). Among cells below direct thresholds of 1%, 2%, and 5%, 13/18, 34/47, and 51/73 remain at or above the same threshold on equal-family indirect ASR. The corresponding rates, 72.2%, 72.3%, and 69.9%, change little across the three operating points; low direct ASR therefore does not become more reliable as the direct threshold relaxes. The association is statistically nonzero, but its magnitude and threshold behavior are inadequate for the stopping decision. The next tests identify what the direct score omits before comparing complete method endpoints.

Table 1: Main-text surface and scheduling summary on Stack A. Attack columns are ASR (%); edit counts are per fact, and random-order ranges are mean ± standard deviation across ten valid orders. Dense and cache/reranker omissions localize complementary pathways; the surface-complete rows compare allocation policies descriptively under matched maximum budgets.

Variant	Para.	Hop	Turn	72h drift	Edits
Full PATHSWEEP	8.2	10.4	12.8	9.4	126
No dense repair	12.9	15.8	13.6	10.8	104
No cache/reranker	8.6	10.9	17.1	14.8	109
No-Graph	9.8	12.7	14.1	10.9	139
Random order	9.5±0.6	12.0±0.8	14.0±0.9	10.6±0.7	151±12
Val.-selected	8.8	11.2	13.3	9.9	134

Table 2: Leakage (%) after freezing each repair. The disjoint bank changes prompt manifests; the independent column changes attacker model, authorship, and attack composition. Its five-family macro includes target-bearing reintroduction. Intervals follow the unit of randomization.

	Disjoint prompt bank				Independent
Method	Direct	Para.	Hop	Turn	5-fam. incl. reintro.
Hard Delete	3.4±0.3	19.8±1.3	24.6±1.6	22.9±1.7	25.8±1.6
Retriever FT	2.1±0.2	13.4±0.9	18.2±1.1	15.9±1.2	19.0±1.3
RAG-UL	2.4±0.3	15.1±1.0	19.3±1.2	17.8±1.3	17.7±1.2
No-Graph	1.3±0.2	11.6±0.8	14.8±1.0	15.1±1.1	16.0±1.1
PATHSWEEP	1.0±0.1	9.1±0.7	11.7±0.9	13.3±1.0	13.2±1.0

6.2 RESIDUAL PATHWAYS LOCALIZE FAILURES AND GUIDE REPAIR

Residual risk is distributed across complementary surfaces. Direct deletion moves direct leakage from 42.1% to 2.4% while leaving 16.7/21.3% paraphrase/multi-hop leakage. Dense repair produces the largest semantic reduction; sparse and metadata edits further suppress alias routes; cache/reranker invalidation yields the largest multi-turn and routine-drift reduction. Leaving dense repair out raises paraphrase/multi-hop leakage to 12.9/15.8%; leaving cache/reranker state intact raises multi-turn/72-hour drift to 17.1/14.8%. The complete surface and ordering ablation appears in Table 9, and target-bearing reintroduction pathways in Table 11.

The matched schedules isolate allocation from surface access. No-Graph nearly matches direct leakage, utility, and runtime, yet is 1.6/2.3/1.3/1.5 points higher on paraphrase, multi-hop, multi-turn, and routine drift. Reverse ordering reaches 9.1/11.6/13.7/10.2% on those endpoints, while ten random valid orders average 9.5±0.6/12.0±0.8/14.0±0.9/10.6±0.7%. The strongest alternative, selected by validation leakage, reaches 8.8/11.2/13.3/9.9%; PATHSWEEP reaches 8.2/10.4/12.8/9.4% with 126 rather than 134 edits per fact. The consistent direction across fixed, reverse, random, and validation-selected controls supports the complete trace-conditioned allocation rule—coverage ranking, retain-loss ranking, and stopping—while the shared surfaces rule out added edit access as the explanation.

6.3 FROZEN EVALUATIONS RETAIN THE RESIDUAL-RISK SEPARATION

The stopping failure persists under both prompt and attacker turnover. Table 2 first changes prompt manifests and then changes attacker model, authorship, and attack composition. Every held-out bank is harder than the core suite, but PATHSWEEP remains lowest on all indirect columns. Under the independent suite, its indirect macro-average is 13.2±1.0, compared with 25.8±1.6, 19.0±1.3, 17.7±1.2, and 16.0±1.1. Attack-family turnover raises absolute leakage and widens intervals while preserving the distinction among stopping decisions.

Blinded validation on 1200 method-stratified outputs yields 94.7% precision, 93.2% recall, 4.8% exact/soft disagreement, and Krippendorff's $\alpha = 0.89$ between adjudicators. These checks support answer-level comparability without relying on the attack generator as judge.

Table 3: Stack-A attack success rate (%, mean ± 95% CI), global retain MRR, and median repair time. Nearly tied direct ASR values correspond to sharply different indirect-risk decisions.

Method	Direct	Para.	Multi-hop	MRR	Time (s)
Hard Delete	2.4±0.3	16.7±1.1	21.3±1.3	0.812	12
Delete+Rebuild	1.8±0.2	14.2±1.0	19.8±1.2	0.815	340
Lexical Deny	1.2±0.2	15.8±1.1	20.1±1.2	0.782	5
Dense Masking	3.1±0.4	13.4±0.9	18.7±1.1	0.808	8
Retriever FT	1.6±0.3	11.2±0.8	15.4±1.0	0.798	180
RAG-UL	1.4±0.2	12.6±0.9	16.1±1.0	0.807	96
No-Graph	0.9±0.1	9.8±0.7	12.7±0.9	0.795	231
PATHSWEEP	0.8±0.1	8.2±0.6	10.4±0.8	0.793	245

Table 4: Component-replaced replication on HotpotQA after repeating the validation procedure under shared cross-stack budgets. Indirect is the macro-average of four query-only families; AnyLeak is the reintroduction-augmented union. Attack metrics are percentages with five-seed 95% intervals; MRR is local retain, and time is the median in seconds. Paired retain intervals appear in Table 16.

Method	Direct	Indirect	AnyLeak	72h drift	MRR	Time
None	44.8±1.8	36.9±1.5	73.4±2.2	43.0±1.9	0.742	0
Hard Delete	2.9±0.4	19.4±1.2	43.9±2.4	24.7±1.5	0.731	14
Retriever FT	1.9±0.3	13.8±0.9	32.4±2.1	18.5±1.2	0.716	196
RAG-UL	1.6±0.2	15.1±1.0	34.7±2.2	20.4±1.3	0.733	105
No-Graph	1.1±0.2	11.5±0.8	27.0±1.9	14.6±1.0	0.708	249
PATHSWEEP	0.9±0.1	9.4±0.7	22.1±1.7	12.3±0.9	0.710	266

6.4 METHOD CHOICES AFTER CRITERION, MECHANISM, AND REPLICATION TESTS

Only after the proxy, surface, and turnover tests do we compare complete endpoints. Table 3 shows the practical consequence: methods that appear nearly tied on direct ASR remain widely separated on paraphrase and multi-hop reconstruction. RAG-UL represents the closest intervention class, while No-Graph provides the surface-matched allocation control.

PATHSWEEP reduces Hard Delete's paraphrase and multi-hop leakage by 8.5 and 10.9 points, respectively, while Hard Delete retains 0.019 more global MRR. RAG-UL preserves stronger utility and costs 149 fewer seconds, but leaves 4.4/5.7 points more paraphrase/multi-hop leakage. Against matched surface access, PATHSWEEP is lower than No-Graph by 1.6/2.3 points and uses fewer edits (Table 1). These comparisons expose distinct operating choices rather than a single monotone ranking.

Component-replaced stack replication. Table 4 changes the task, dense and sparse retrievers, reranker, generator family, and cache policy while repeating the validation procedure under shared cross-stack budgets. Direct leakage again compresses while indirect macro-risk and reintroduction-augmented fact compromise remain separated. The residual profile identifies PATHSWEEP as the lowest-risk retrieval-side repair, whereas RAG-UL provides the strongest repaired utility/time point. The surface-matched gap has the same direction on both stacks: AnyLeak@32 is 18.7% versus 22.4% on Stack A and 22.1% versus 27.0% on Stack B for PATHSWEEP and No-Graph, respectively.

Operating choices after risk is measured. The residual-risk gains remain meaningful after accounting for neighboring knowledge. On Stack A, PATHSWEEP changes local MRR from 0.821 to 0.789 and global MRR from 0.815 to 0.793; its nearby Recall@10 is 0.843 [0.832,0.854], answer exact match is 0.652 [0.637,0.667], and 5.2% [4.3,6.2] of previously correct nearby answers are lost. Hard Delete loses fewer correct answers (3.7% [3.0,4.5]) but leaves twice the fact compromise. PATHSWEEP achieves lower indirect risk than Retriever FT and No-Graph at overlapping paired utility; RAG-UL remains the strongest utility/cost point. Structural slices and the full Pareto analysis are reported in Tables 17–18.

Table 5: Fact compromise and benign-utility frontier on Stack A. AnyLeak@32 is the reintroduction-augmented union; it and previously correct answers lost are percentages. All ranges are 95% fact- or query-clustered intervals. Time is the median complete repair work item in seconds.

Method	AnyLeak@32	Local MRR	Correct lost	Time
Hard Delete	38.6% [36.2,41.1]	0.801±0.006	3.7% [3.0,4.5]	12
Retriever FT	27.3% [25.1,29.7]	0.794±0.007	4.9% [4.0,5.9]	180
RAG-UL	29.1% [26.9,31.5]	0.805±0.006	3.2% [2.6,3.9]	96
No-Graph	22.4% [20.3,24.6]	0.787±0.007	5.6% [4.7,6.7]	231
PATHSWEEP	18.7% [16.4,21.2]	0.789±0.007	5.2% [4.3,6.2]	245

7 DISCUSSION

The statistical and systems results support the same decision. Direct leakage is a useful channel measurement, but its 0.23 rank correlation, 0.61 AUROC, and roughly 70% false-safe rate make it an insufficient summary after direct responses approach the floor. Query-only channel rates address black-box reconstruction without ingestion; reintroduction-augmented AnyLeak@32 addresses the stronger capability to submit target-bearing content. Paired benign utility and runtime then determine which risk reductions are operationally worthwhile.

The trace experiments explain the lost information. Exact deletion closes the shortest route, while semantic neighbors, sparse aliases, metadata links, reranker features, and cached derivations preserve attack-specific alternatives. Leave-one-surface tests connect dense state to semantic reconstruction and cache/reranker state to temporal reconstruction. No-Graph, random order, and the validation-selected schedule hold the editable surfaces and maximum budgets fixed; PATHSWEEP's lower leakage with fewer edits is consistent across the four displayed endpoints, independent attacks, and both stacks. Together, trace attribution, surface omissions, and matched schedules identify allocation—rather than access to extra edit operations—as the supported mechanism.

The evidence applies to baseline-answerable protected relations in two English QA stacks where dense, sparse, metadata, reranker, and cache state are observable; multilingual aliases, routine ingestion, and a retrieval/LM factorial broaden the tested routes and substrates. Independent attacker turnover is evaluated on Stack A, while Stack B isolates task and component replacement under the core protocol. Repair probes and final banks share the answer judge but use disjoint manifests; independent prompt authors and blinded adjudication strengthen comparisons within this declared evaluation. The black-box query results measure empirical reconstruction, while target-bearing reintroduction deliberately evaluates a stronger ingestion capability. These endpoints quantify compromise under the declared attacks rather than a cryptographic deletion or universal privacy guarantee. Within this regime, PATHSWEEP spends additional time and local MRR to minimize retrieval-side compromise, RAG-UL supplies a stronger utility/cost point, and the combined retrieval/LM repair reaches lower leakage with 9.4% correct-answer loss and 665 seconds. The threshold-sweep hypervolume favors PATHSWEEP (0.78 versus 0.74 for combined), so policy-specific risk, utility, and cost constraints can select an operating point without collapsing them into one score.

8 CONCLUSION

Forgetting decisions in hybrid RAG must account for the execution paths that can still reconstruct a protected relation. RRA-BENCH makes that state measurable through a fact-level residual-memory profile, separated query-only and reintroduction threat models, and proxy-validity statistics that preserve the crossed repair design. The 0.23 rank correlation, 0.61 fact-compromise AUROC, and stable 69.9–72.3% false-safe rates show that a direct-answer pass cannot carry the stopping decision alone. Surface and scheduling experiments identify where that missing risk resides and demonstrate how traces can guide multi-surface repair under explicit utility and cost constraints. Residual retrieval memory therefore supplies both the scientific object and the practical decision layer for evaluating forgetting in stateful RAG systems.

AI USE STATEMENT

Generative AI tools assisted manuscript preparation, including drafting and revising prose; title, abstract, frontmatter, and organizational framing; related-work synthesis and reference formatting; and formal, methodological, and experimental framing. The available preparation record does not establish that generative AI produced the underlying measurements or independently ran the reported experiments.

REPRODUCIBILITY STATEMENT

Sections 3–6 specify the protected object, attacker capabilities, trace construction, repair algorithm, budgets, stacks, comparators, randomization, adjudication, statistics, hardware, runtime work item, and primary results. The appendices provide secondary results, implementation settings, uncertainty conventions, and a systems-and-security matrix mapping each conclusion to its experimental variation.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

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

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

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, pp. 2633–2650, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. doi: 10.48550/arXiv.2312.10997. URL https://arxiv.org/abs/2312.10997.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Y. Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. 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 *International Conference on Machine Learning*, pp. 3832–3842, 2020.

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

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 874–880, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021. doi: 10.48550/arXiv.2112.09118. URL https://arxiv.org/abs/2112.09118.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023. URL https://www.jmlr.org/papers/v24/23-0037.html.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 14389–14408, 2023. doi: 10.18653/v1/2023.acl-long.805.

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

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020.

Junbeom Kim, Kyuyoung Kim, Jihoon Tack, Dongha Lim, and Jinwoo Shin. Scalable and robust LLM unlearning by correcting responses with retrieved exclusions, 2025. URL https://arxiv.org/abs/2509.25973.

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/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2356–2362, 2021. doi: 10.1145/3404835.3463238. URL https://doi.org/10.1145/3404835.3463238.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. *arXiv preprint arXiv:2401.06121*, 2024.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Yuefeng Peng, Parnian Afshar, Megan Ganji, Thomas Butler, Amir Houmansadr, Mingxian Wang, and Dezhi Hong. Forget to know, remember to use: Context-aware unlearning for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://icml.cc/virtual/2026/poster/60740.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 2523–2544, 2021. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200/.

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

Raj Sanjay Shah, Jing Huang, Keerthiram Murugesan, Nathalie Baracaldo, and Diyi Yang. The unlearning mirage: A dynamic framework for evaluating LLM unlearning. In *Second Conference on Language Modeling*, 2025. doi: 10.48550/arXiv.2603.11266. URL https://arxiv.org/abs/2603.11266.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, 2025.

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. ACM, 2026. doi: 10.1145/3805712.3808545. URL https://doi.org/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 *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Kuan Wang and Chao Zhang. MemLeak: Diagnosing information leaks in multimodal agent memory, 2026. URL https://arxiv.org/abs/2606.29788.

Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. When machine unlearning meets retrieval-augmented generation (RAG): Keep secret or forget knowledge? *IEEE Transactions on Dependable and Secure Computing*, 23(2):2030–2045, 2026. doi: 10.1109/TDSC.2025.3620832. URL https://doi.org/10.1109/TDSC.2025.3620832.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tacl_a_00595. URL https://aclanthology.org/2023.tacl-1.63/.

A Implementation and Evaluation Details

A.1 Cohort Construction and Evaluation Partitions

The construction pipeline is identical across stacks. We first require baseline retrieval of a supporting passage and a correct baseline answer, then align a canonical target relation to its answer spans, aliases, entities, and supporting documents. We remove examples without an adjudicable entity–relation pair, collapse normalized duplicate triples and duplicate sources, and exclude cases where an unmarked alternative passage states the target directly. The final sample contains 500 protected facts per stack. Alias maps, source alignments, and relation labels are frozen before prompt authoring.

Nearby retention is constructed after target selection. The four mutually exclusive slices in Table 18 contain 420 graph-distance-one, 530 dense-similarity-only, 460 entity-overlap-only, and 590 relation-overlap-only queries. Every nearby query is baseline answerable and excludes the complete protected triple and its answer aliases; relation-overlap queries share only the predicate. Global retention draws 5,000 baseline-answerable queries after removing target entity, relation, answer, and source-document overlap. Validation and reported retention manifests are query-disjoint. Repair probes, core scoring prompts, disjoint prompts, and independent attacks use the same fact identities but non-overlapping prompt strings and authoring manifests.

A.2 Criterion-Validity Estimator

The 150-cell grid consists of six repair initializers, five cumulative surface checkpoints, and five seeds. The post-initializer checkpoint records each method's native repair; later checkpoints use the common surface operators, with non-PathSweep candidates ordered by estimated retain loss and seed-fixed identifier and PathSweep candidates ordered by $s(e)$. For each cell, rank association compares macro direct ASR with a channel-equal macro-average over five indirect families: paraphrase, multi-hop, multi-turn, multilingual alias, and target-independent 72-hour drift. The AUROC expands each cell to its 500 fact observations: the predictor is per-fact direct ASR and the binary outcome is success anywhere in the frozen, reintroduction-augmented AnyLeak@32 manifest. A bootstrap replicate samples fact identifiers with replacement and carries along every method, checkpoint, and seed value for that fact before recomputing the complete statistic. This crossed resampling yields the intervals in Figure 1 without counting 75,000 fact–cell pairs as independent evidence.

The AnyLeak manifest allocates its 32 endpoint queries as eight paraphrase, eight multi-hop, six multi-turn, six multilingual-alias, and four controlled-reintroduction positions. Direct prompts are excluded. Removing the four intervention positions yields the query-only $\mathrm{AnyLeak@28}_Q$ submanifest; all AnyLeak@32 tables report the stronger $\mathrm{AnyLeak@32}_R$ union. Each reintroduction is evaluated from a fresh post-repair snapshot and discarded after its endpoint query; ingestion actions do not consume endpoint-query positions. All positions are evaluated; AnyLeak records their union over each fact–seed pair, and median queries-to-first-leak is conditional on compromised pairs. Identical seed-specific manifests are replayed across methods, and intervals resample facts while preserving all five outcomes for each fact.

A.3 Comparator and Repair Settings

Hard Delete removes every aligned source chunk from dense and sparse stores. Delete+Rebuild applies the same removal and rebuilds FAISS rather than mutating it in place. Lexical Deny normalizes the protected answer and aliases with the benchmark normalizer and filters matching candidates before reranking. Dense Masking sets the cosine scores of aligned normalized-document IDs to $-\infty$ before candidate fusion, avoiding an undefined zero-vector normalization. Retriever FT applies the pairwise repulsion loss for 500 updates at learning rate 10^{-5} with negatives drawn from the retain-validation neighborhood; its edits are rejected by the common 0.02 absolute-MRR guardrail.

RAG-UL follows the published retrieval-score objective and constraint-context construction (Wang et al., 2026); its retrieval and constraint losses are optimized jointly on the repair split, with validation retention selecting the checkpoint under the common 0.02 MRR guardrail. The outer facts, retrieved candidates, six repair-probe rounds, evaluator, seeds, and stopping tests are shared with the other retrieval-side methods. No-Graph invokes the same direct, dense, sparse, metadata, reranker, and cache candidate generators as PathSweep, but sorts them by the fixed global surface order before

Table 6: Stack-A fact compromise under the 32-query reintroduction-augmented budget. Counts pool 500 facts across five seeds as 2,500 fact–seed pairs; intervals cluster by fact and retain its five outcomes.

Method	AnyLeak@32	Compromised / 2500	Median queries
Hard Delete	38.6% [36.2,41.1]	965/2500	7
Delete+Rebuild	34.8% [32.4,37.3]	870/2500	8
Lexical Deny	37.4% [35.0,39.9]	935/2500	7
Dense Masking	33.1% [30.8,35.5]	828/2500	9
Retriever FT	27.3% [25.1,29.7]	683/2500	11
RAG-UL	29.1% [26.9,31.5]	728/2500	10
No-Graph	22.4% [20.3,24.6]	560/2500	14
PATHSWEEP	18.7% [16.4,21.2]	468/2500	17

the seed-fixed candidate identifier. Random-order controls uniformly permute precedence-respecting candidate orders; the validation heuristic chooses the order with lowest validation indirect ASR under the same retain guardrail. PATHSWEEP instead recomputes active-path coverage and counterfactual validation MRR after every committed edit. These controls vary the complete allocation policy while holding editable state, candidate generation, maximum applied edits, probe rounds, and termination tests fixed.

For the retrieval/LM factorial, LM-only maximizes target-token loss on the protected answer spans, context-aware LM uses the published context-conditioned retain objective (Peng et al., 2026), and combined repair applies PATHSWEEP before the context-aware LM update. All three use the same protected facts and validation partition; target loss is evaluated on the canonical answer and frozen aliases, retain regularization uses the nearby-query validation set, and checkpoint selection minimizes validation leakage subject to the reported utility guardrail. Final attack prompts and reported retain queries never enter optimization or stopping.

A.4 SYSTEM IMPLEMENTATION AND ADJUDICATION

Implementation. Stack A starts from the released Contriever-MSMARCO encoder, Pyserini BM25, MiniLM-L6-v2, and FAISS IVF4096,PQ32; Retriever FT and admitted dense-repair operations update separate cloned encoders for 500 steps at learning rate 10^{-5} under the repulsion loss. Stack B uses BGE-M3 dense vectors, SPLADE-v3 postings, bge-reranker-large, and HNSW32. Every protected fact is repaired and evaluated in an isolated system clone initialized from the same pre-repair image, so edits and ingestion never accumulate across facts. Llama-3.1-70B and Mistral-Large-Instruct-2407 decode with temperature 0.2, top-p 0.9, and 128 maximum new tokens. Total training and evaluation cost is 380 A100-GPU-hours.

Adjudication. Attack generation never receives answer-judge instructions or labels. Exact normalization lowercases, strips punctuation and articles, and resolves aliases through the frozen per-fact map. The soft rule requires matching target entity and relation and rejects topical, entity-swapped, and relation-broadened answers. If exact and soft decisions differ, the union defines ASR and the pair is retained for disagreement analysis. Calibration samples are balanced over method and family; disagreements are resolved by a third blind adjudicator. Across method-by-family cells, precision ranges from 92.1–96.8% and recall from 90.4–95.9%. The largest false-negative rate is 8.6% for Hard Delete under target-bearing reintroduction probes, and the largest false-positive rate is 6.2% for RAG-UL multi-hop outputs.

Core-bank denominator ledger. Each Stack-A seed contains 5000 direct, 10000 paraphrase, and 10000 multi-hop attempts. The following seed-equivalent ratios multiply the five-seed mean by one seed's denominator; they expose the rate denominator but are not treated as pooled binomial samples. Hard Delete yields 120/5000, 1670/10000, and 2130/10000; PATHSWEEP yields 40/5000, 820/10000, and 1040/10000.

Table 7: Complete Stack-A profile for the closest and surface-matched comparators. Attack columns are percentages with five-seed 95% intervals.

Method	Multi-turn	72h drift	Edits / fact
RAG-UL	16.9±1.1	15.8±1.1	82
No-Graph	14.1±0.9	10.9±0.8	139
PATHSWEEP	12.8±0.9	9.4±0.7	126

Table 8: Paired utility and synchronized runtime for the closest and surface-matched comparators. Times are median [IQR] seconds over seven repetitions.

Method	Local MRR	Global MRR	Time [IQR]
RAG-UL	0.805±0.006	0.807±0.004	96 [90,104]
No-Graph	0.787±0.007	0.795±0.004	231 [219,246]
PATHSWEEP	0.789±0.007	0.793±0.004	245 [232,261]

Table 6 exposes the fact-level consequence hidden by aggregate direct ASR. The ordering agrees with the indirect channel rates, while the median successful position among compromised fact–seed pairs shows that stronger repairs require more endpoint queries before compromise.

The multi-turn and routine-drift columns preserve the same separation among the closest intervention and scheduling controls. PATHSWEEP also applies fewer edits than No-Graph, linking its risk advantage to allocation rather than a larger edit count.

RAG-UL is the fastest and highest-retention point in this matched subset. No-Graph and PATHSWEEP have overlapping retention, so their leakage gap is interpreted together with surface access and edit-count controls rather than as a utility advantage.

B STAGE, DRIFT, AND REINTRODUCTION DIAGNOSTICS

This section separates three mechanisms that are aggregated in the main decision tables: cumulative repair depth, benign ingestion over time, and controlled target-bearing reintroduction. The displays use the same protected facts and report attack success at the endpoint of each intervention.

Routine ingestion contains ordinary corpus updates sampled independently of the target and replays the same seed-specific stream for every method. The post-repair state is measured at 0, 6, 12, 24, 48, and 72 hours. Updates enter through each stack's standard ingestion path, which refreshes affected dense and sparse index state, reranker inputs, and summary-cache state; the evaluation never inserts a

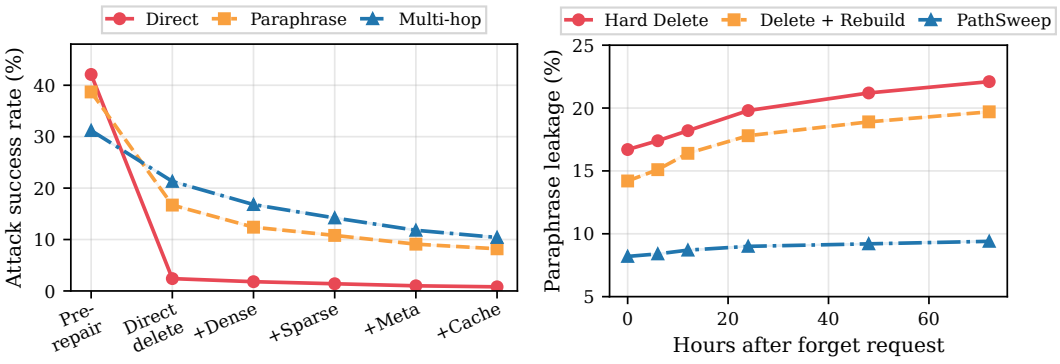


Figure 2: Stack-A stagewise attack success (left) and paraphrase leakage under target-independent routine ingestion (right), shown as five-seed means. Direct deletion closes the shortest path, whereas later surface operations reduce indirect leakage; for PATHSWEEP, paraphrase ASR rises from 8.2% immediately after repair to 9.4% at 72 hours.

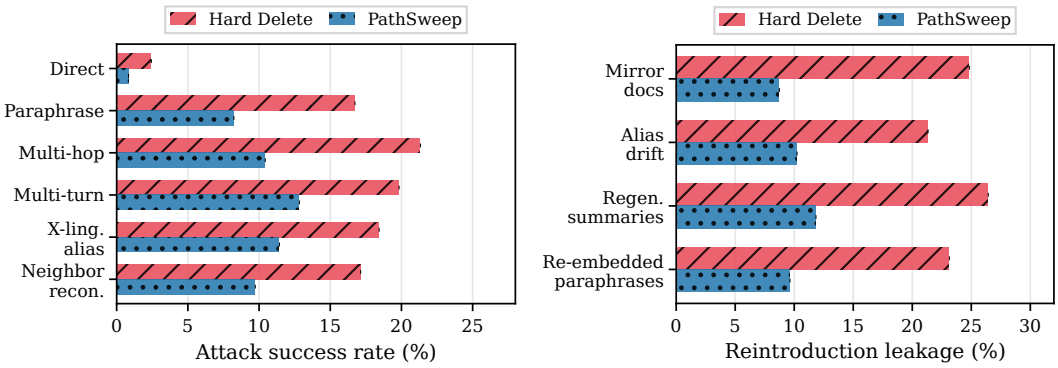


Figure 3: Attack-family profile (left) and explicit target-bearing reintroduction pathways (right), shown as five-seed means. Each right-panel intervention supplies one transformed route to the protected relation before the endpoint probe.

Table 9: Surface and ordering ablations on Stack A. Entries are ASR (%); edit counts are per fact. All variants share edit, retain, stopping, and runtime budgets.

Variant	Direct	Para.	Hop	Turn	72h drift	Edits
Full PATHSWEEP	0.8	8.2	10.4	12.8	9.4	126
No-Graph	0.9	9.8	12.7	14.1	10.9	139
No dense repair	1.3	12.9	15.8	13.6	10.8	104
No sparse repair	1.5	9.6	12.3	13.5	10.5	113
No metadata guard	1.4	9.4	11.9	13.8	10.7	117
No cache/reranker	0.9	8.6	10.9	17.1	14.8	109
Reverse order	1.0	9.1	11.6	13.7	10.2	142
Random order	1.1±0.2	9.5±0.6	12.0±0.8	14.0±0.9	10.6±0.7	151±12
Val.-selected heuristic	0.9	8.8	11.2	13.3	9.9	134

transformed target item into this stream. Thus the curve measures drift under ordinary state refresh, whereas the following snapshot-isolated interventions measure deliberate reintroduction.

Controlled reintroduction answers a different question. Mirror documents restate the source, alias drift changes entity surface forms, regenerated summaries derive a new statement from the ingested item, and re-embedded paraphrases alter the dense representation. These inputs remain inside the forget request's protected relation and may trigger sparse, metadata, cache, or dense guards.

The leave-one-surface rows localize complementary effects rather than one universally dominant component. Dense omission causes the largest semantic increase; cache/reranker omission causes the largest multi-turn and routine-drift increase; fixed and random schedules remain between the strongest omissions and full PATHSWEEP.

Multilingual-alias leakage is most sensitive to sparse and metadata operations, whereas the utility columns stay close across the surface-complete schedules. This combination supports attack-family-specific allocation rather than indiscriminate neighborhood removal.

Regenerated summaries are the strongest controlled Hard Delete pathway at 26.4%; PATHSWEEP reduces that endpoint to 11.8%. The pathway table therefore complements the routine-ingestion curve by identifying which transformations recreate target-bearing evidence.

C RETRIEVAL AND GENERATOR FACTORIAL

The factorial distinguishes retrieval-mediated reconstruction from generator-side memory under one fixed Stack-A interface. Attack and utility tables must be read jointly because the combined intervention reduces leakage further while paying the largest local-retention cost.

Table 10: Alias leakage, utility, and runtime for the surface and ordering ablations. Utility uses paired query-clustered intervals; runtime covers the complete repair work item.

Variant	X-ling. ASR	Local MRR	Global MRR	Time (s)
Full PATHSWEEP	11.4	0.789	0.793	245
No-Graph	13.2	0.787	0.795	231
No dense repair	13.9	0.797	0.799	214
No sparse repair	16.2	0.792	0.796	226
No metadata guard	17.4	0.793	0.797	224
No cache/reranker	12.8	0.791	0.794	218
Reverse order	12.7	0.788	0.792	252
Random order	13.4±0.9	0.786±0.007	0.791±0.004	258±18
Val.-selected heuristic	12.1	0.788	0.793	247

Table 11: Leakage after snapshot-isolated target-bearing reintroduction. Entries are five-seed means with Student-t 95% intervals; each pathway starts from the same post-repair state and is discarded after its endpoint query.

Pathway	Hard Delete	PATHSWEEP
Mirror documents	24.8±1.5	8.7±0.7
Alias drift	21.3±1.3	10.2±0.8
Regenerated summaries	26.4±1.7	11.8±0.9
Re-embedded paraphrases	23.1±1.4	9.6±0.8

Retrieval-only repair explains most of the reduction because it removes the external evidence paths available to an unchanged generator. A context-aware LM baseline following the utility-preserving objective of Peng et al. (2026) improves retain quality over gradient ascent while leaving 13.9–16.7% leakage across the four channels. Combined repair further lowers every attack rate at the cost of another 0.027 local MRR and 420 seconds relative to retrieval-only. The 0.772 and 0.771 global MRR values for combined and LM-only overlap; their consequential difference is leakage.

D INDEPENDENT ATTACKER DETAIL

The independent suite is confirmatory: systems are frozen before Qwen-generated and human-authored prompts are revealed, and authors and judges are blind to method identity. It changes prompt source, attacker model, and attack composition simultaneously.

Independent direct leakage remains the easiest channel at 4.1±0.4, 2.7±0.3, 2.2±0.3, 1.7±0.2, and 1.4±0.2 for Hard Delete, Retriever FT, RAG-UL, No-Graph, and PATHSWEEP, respectively.

AnyLeak@32 is 47.3% [44.6,50.0], 36.6% [34.0,39.3], 34.2% [31.7,36.9], 30.1% [27.7,32.7], and 24.8% [22.5,27.3]. Median successful-query budgets are 6, 8, 9, 11, and 14, respectively. Independent attacks raise every method's absolute risk, but PATHSWEEP's advantage over the same-surface control remains visible in all five families and fact compromise.

Table 12: Retrieval/LM factorial attack rates on Stack A. Entries are ASR means with five-seed 95% intervals.

Repair	Direct	Para.	Hop	Turn
None	42.1±1.8	38.7±1.6	31.2±1.5	28.4±1.4
Retrieval-only	0.8±0.1	8.2±0.6	10.4±0.8	12.8±0.9
LM-only	14.7±1.0	18.1±1.2	16.8±1.1	15.2±1.0
Context-aware LM	13.9±0.9	16.7±1.1	15.9±1.0	14.5±1.0
Combined	0.6±0.1	6.1±0.5	7.8±0.6	9.3±0.7

Table 13: Utility and runtime for the Stack-A factorial. MRR intervals use query-clustered bootstrap.

Repair	Local MRR	Global MRR	Time (s)
None	0.821±0.006	0.815±0.004	0
Retrieval-only	0.789±0.007	0.793±0.004	245
LM-only	0.768±0.008	0.771±0.005	420
Context-aware LM	0.790±0.007	0.801±0.004	448
Combined	0.762±0.008	0.772±0.005	665

Table 14: Independent Qwen/human attack suite after systems are frozen. ASR is percent; each cell is a five-seed mean with a 95% interval. Reintro. is the target-bearing intervention included in Table 2's independent five-family macro.

Method	Para.	Hop	Turn	X-ling.	Reintro.
Hard Delete	22.8±1.5	27.9±1.8	25.4±1.7	23.1±1.5	29.6±1.9
Retriever FT	15.9±1.1	20.7±1.4	18.8±1.3	17.3±1.2	22.1±1.5
RAG-UL	14.8±1.0	19.0±1.3	18.1±1.2	16.5±1.1	20.2±1.4
No-Graph	13.5±0.9	16.9±1.2	17.2±1.2	15.8±1.1	16.4±1.1
PATHSWEEP	10.8±0.8	13.6±1.0	14.9±1.1	12.7±0.9	14.2±1.0

E SECOND-STACK DETAIL

Stack B uses 5000 direct, 10000 paraphrase, 10000 multi-hop, 6000 multi-turn, and 4000 multilingual-alias prompts per seed. Table 15 exposes the family values behind the main paper's indirect macro-average.

All four query-only families favor the same residual-risk endpoint even though their absolute margins differ from Stack A. Repeating the validation procedure with shared budgets tests the criterion under a different retrieval geometry rather than repeating the first stack with new prompts.

RAG-UL retains the strongest local utility and runtime among repaired multi-operation systems, while No-Graph and PATHSWEEP occupy lower-risk points. This reproduces the non-monotone choice among risk, retention, and cost.

F COLLATERAL DAMAGE AND PARETO ANALYSIS

This analysis asks whether lower residual risk is purchased by deleting broad neighborhoods around each protected fact. It evaluates the same fixed nearby queries before and after repair and preserves query-to-fact pairing in the intervals.

Among 1368 previously correct nearby answers, losses are 51, 67, 44, 77, 71, 114, and 129 for Hard Delete, Retriever FT, RAG-UL, No-Graph, PATHSWEEP, LM-only, and combined repair. Median absolute rank displacements are 1.8, 2.4, 1.6, 2.7, 2.5, 3.1, and 3.5, respectively.

Table 15: Stack-B attack success by family (%, five-seed mean ± 95% interval).

Method	Direct	Para.	Hop	Turn	X-ling.
None	44.8±1.8	39.2±1.7	38.5±1.6	35.4±1.5	34.5±1.5
Hard Delete	2.9±0.4	20.6±1.3	21.7±1.4	17.9±1.2	17.4±1.2
Retriever FT	1.9±0.3	14.2±1.0	15.5±1.1	13.4±0.9	12.1±0.9
RAG-UL	1.6±0.2	15.0±1.0	17.2±1.2	14.4±1.0	13.8±1.0
No-Graph	1.1±0.2	11.1±0.8	12.9±0.9	11.8±0.8	10.2±0.8
PATHSWEEP	0.9±0.1	8.8±0.7	10.7±0.8	9.7±0.7	8.4±0.6

Table 16: Stack-B local/global utility with paired query-clustered 95% intervals and synchronized median runtime.

Method	Local MRR	Global MRR	Time (s)
None	0.742±0.006	0.736±0.004	0
Hard Delete	0.731±0.006	0.737±0.004	14
Retriever FT	0.716±0.007	0.724±0.005	196
RAG-UL	0.733±0.006	0.733±0.004	105
No-Graph	0.708±0.007	0.719±0.005	249
PATHSWEEP	0.710±0.007	0.718±0.005	266

Table 17: Stack-A nearby-query utility. MRR and correct-answer-loss ranges are paired query-clustered 95% intervals; the remaining retention metrics use the same fixed query set.

Method	MRR	Recall@10	Answer EM	Top-k overlap	Correct lost
None	0.821±0.006	0.873	0.684	1.000	0.0%
Hard Delete	0.801±0.006	0.857	0.663	0.946	3.7% [3.0,4.5]
Retriever FT	0.794±0.007	0.848	0.654	0.928	4.9% [4.0,5.9]
RAG-UL	0.805±0.006	0.861	0.667	0.951	3.2% [2.6,3.9]
No-Graph	0.787±0.007	0.840	0.648	0.919	5.6% [4.7,6.7]
PATHSWEEP	0.789±0.007	0.843	0.652	0.924	5.2% [4.3,6.2]
LM-only	0.768±0.008	0.831	0.627	0.907	8.3% [7.1,9.7]
Combined	0.762±0.008	0.825	0.620	0.899	9.4% [8.1,10.9]

Sweeping repair thresholds yields normalized leakage–utility hypervolume 0.78 for PATHSWEEP, 0.72 for No-Graph, 0.69 for RAG-UL, 0.61 for Retriever FT, 0.53 for Hard Delete, and 0.74 for combined retrieval/LM repair. Leakage and local-MRR axes are min–max normalized over the observed threshold sweep, and dominated area is measured from the worst observed leakage/utility corner. The combined method reaches lower leakage at larger local damage; PATHSWEEP has the largest dominated area over this operating range.

The four disjoint slices show that utility loss is concentrated near the protected relation. PATHSWEEP improves on No-Graph in each slice, including the graph-distance-one cases that are most exposed to collateral edits.

Table 18: Disjoint nearby-query slices on Stack A. The final columns report post-repair MRR and previously correct answers lost.

Slice	n	Base MRR	No-Graph	PATHSWEEP	NG lost	PS lost
Graph distance 1	420	0.806	0.761	0.766	7.2%	6.5%
Dense-similarity only	530	0.818	0.779	0.782	6.1%	5.7%
Entity overlap only	460	0.829	0.790	0.795	5.5%	4.8%
Relation overlap only	590	0.830	0.806	0.811	3.9%	3.5%

G Systems and Security Evidence Matrix

The matrix summarizes which conclusion each experiment identifies by separating the protected object from the attacker or intervention varied in that test.

Table 19: Evaluation boundary and threat coverage.

Setting	Protected object	Test variation	Target
Core	Fact relation	In-family probes	Proxy validity
Disjoint	Fact relation	New manifests	Prompt transfer
Independent	Fact relation	Qwen/human probes	Attacker transfer
Surface control	Fact relation	Complete allocation rule	Allocation value
Surface ablation	Fact relation	Leave one surface out	Attribution
Routine drift	Fact relation	Target-independent updates	State-refresh stability
Reintroduction	Fact relation	Target-bearing ingestion	Ingestion attack
LM factorial	Fact relation	Retrieval vs. LM repair	Substrate attribution
Stack B	Fact relation	Task/components replaced	Protocol replication

The matrix distinguishes prompt transfer, attacker-family transfer, component replication, and substrate attribution. Core and surface studies adapt within a fixed stack to isolate mechanism; disjoint and independent studies freeze repair before prompt revelation. Across these settings, indirect evidence paths remain the repeated failure after direct suppression, while the dominant route changes with the intervention and the reported utility and runtime expose the cost of closing it.