VERSIONSEAL: EXACT-VERSION AUTHORIZATION FROM FEDERATED ROUTE COMMIT TO CITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two federated retrievals can finish with the same selected site, source identifier, and cited key-version yet disagree on whether that citation was authorized. If a replacement version was visible when routing inspected the site, the old version violates the route snapshot; if publication occurred immediately afterward, the identical citation is valid. Final freshness and provenance records cannot distinguish these histories. VERSIONSEAL preserves the missing witness by atomically naming each query-visible index generation, sealing the selected generations at route commit, and judging every version-addressed cache path against that token before service in Strict mode or asynchronously in Balanced mode. On three paired six-hour replays, the complete Balanced protocol reduces exact-snapshot violations from 2.12% for FreshSketch—a same-stack freshness-aware control without atomic generation publication, route sealing, or exact-generation path disposition—to 1.18% (paired reduction 0.94 pp [0.78,1.09]); p99 retrieval latency changes from 4.84 to 4.62 s and WAN traffic from 146 to 133 KB/query. Strict operation records 0 violations among 443700 served paths at 98.6% query availability, while Balanced provides 99.2% availability and, with generator and context budget fixed, improves temporal-answer correctness by +4.6 pp [1.9,7.2] and reduces outdated or unsafe answers by -3.9 pp [-6.1,-1.8]. A 14-day Wikidata/Wikipedia revision replay and ten paired 24-hour runs preserve the complete-protocol violation-rate ordering; together with the observational separation, these results support treating route commit as an authorization boundary under the evaluated update dynamics.

1 INTRODUCTION

Consider the same FDA dosage citation at the end of two executions. In history H_1, routing observes generation $G^{(2)}$, which already contains replacement dosage v_2, but an asynchronous path cache returns v_1. In history H_2, routing observes $G^{(1)}$ and the publication of v_2 occurs immediately afterward, so the same cached v_1 is authorized. Once both executions finish, their final site frontier, selected site, cited source, and cited version are identical. Only the generation visible when the route committed separates a violation from a valid citation.

This is not a corner case of source attribution. A freshness check answers whether a site may compete, and provenance identifies where a returned claim came from; neither states whether the exact returned version belonged to the visibility state that authorized that site. In the primary replay, FreshSketch combines route-visible watermarks, pre-score eligibility, selective fanout, and version-addressed paths, yet 2.12% of its returned paths disagree with the route snapshot. A generator can therefore receive policy-inconsistent evidence even when the selected site and cited source both look correct.

Temporal retrieval and dynamic-knowledge benchmarks make recent evidence easier to find (Vu et al., 2024; Wu et al., 2024; Ouyang et al., 2025; Tang et al., 2025); federated search chooses which repositories to query (Callan et al., 1995; Kulkarni & Callan, 2015; Wang et al., 2024; Guerraoui et al., 2025); and citation methods expose supporting material (Gao et al., 2023; Stolfo, 2024). These capabilities treat H_1 and H_2 alike unless the generation on which they differ is carried from route commit through cache service and merge. Storage snapshots and watermarks can name that generation; VERSIONSEAL makes it the request-scoped reference for checking each returned key-version.

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.

VERSIONSEAL realizes the join as an ordered route-sealing protocol. At gate-snapshot time t_g, the coordinator reads each site's atomically published watermark and immutable generation, removes sites outside the query's freshness budget, and scores only the survivors. Route commit at t_r seals the selected sites and generations. A site may answer from its asynchronous path cache, whose entries carry exact key-versions but are not keyed by the route token; the named generation therefore supplies the authority against which the cache payload is checked. Strict mode completes that comparison before service and otherwise reseals, retries, or abstains. Balanced mode launches the same comparison but may merge a deadline-ready cache path while validation is outstanding, exposing a measured availability–enforcement operating point rather than changing the predicate.

The complete Balanced protocol reduces the violation rate from 2.12% to 1.18% relative to same-stack FreshSketch, a paired reduction of 0.94 pp [0.78,1.09]; its 4.62 s p99 latency and 133 KB/query traffic are also lower than FreshSketch's 4.84 s and 146 KB/query in those runs. Strict mode records 0 violations among 443700 served paths at 98.6% availability. With questions, generator, decoding, and context budget fixed, Balanced VERSIONSEAL produces 998/1200 (83.2%) temporally correct answers versus 943/1200 (78.6%) for FreshSketch. The ordering persists in a 14-day public revision replay and ten paired 24-hour runs.

The argument follows the distinguishing generation. Section 3 formalizes the H_1/H_2 separation and a route-admissible path predicate. Section 4 implements that predicate as pre-service enforcement in Strict mode and a recorded asynchronous disposition in Balanced mode. Section 5 tests the complete protocol, both dispositions, each observable's role, and the downstream temporal-QA consequence. Together, these results provide an executable exact-version authorization interface and a measured enforcement–availability frontier for mutable federated retrieval.

2 RELATED WORK

Dynamic and version-aware RAG addresses which historical state a question requests. FreshLLMs and TS-Retriever target current or time-sensitive retrieval (Vu et al., 2024; Wu et al., 2024); MRAG separates semantic and temporal relevance (Zhang et al., 2025); VersionRAG explicitly represents document-version evolution (Huwiler et al., 2025); and HoH and EvoWiki evaluate outdated or evolving knowledge (Ouyang et al., 2025; Tang et al., 2025). These methods motivate exact temporal evidence, but their query-time target is different from the execution-time question here: after a site is admitted, did every returned version belong to the precise generation consumed by that routing decision?

Federated retrieval decides where to search. Classical selective search estimates useful collections (Callan et al., 1995; Si & Callan, 2003; Kulkarni & Callan, 2015); FeB4RAG evaluates federated search inside a RAG pipeline (Wang et al., 2024); and RAGRoute learns low-overhead source selection (Guerraoui et al., 2025). Their routing objectives can be used inside VERSIONSEAL, which leaves the relevance scorer unchanged. The additional object is an execution contract joining source selection to mutable, version-addressed path service. FreshSketch is the controlled same-stack comparator for this distinction: it has watermarks, pre-score eligibility, and returned version IDs, but no request-scoped generation authority over cache paths.

Database and streaming mechanisms supply the necessary state semantics. Multi-version stores and snapshot reads name a consistent storage view (DeCandia et al., 2007; Corbett et al., 2012; Lloyd et al., 2011); logical clocks, bounded-staleness estimates, and watermarks expose order or progress (Lamport, 1978; Bailis et al., 2012; Akidau et al., 2013; 2015; Carbone et al., 2015). A storage read timestamp governs objects read within that transaction, whereas federated retrieval first consumes per-site metadata to choose sites and later merges independently materialized cache paths. VERSIONSEAL does not replace snapshot isolation: it turns the selected snapshots into a route-level authorization token and verifies each generator-bound path against the corresponding named generation.

Citation and groundedness methods identify evidence after retrieval (Gao et al., 2023; Stolfo, 2024; Fang et al., 2024); RAG, REALM, RETRO, Atlas, and active or self-reflective retrieval improve content selection and retrieval–generation interaction (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023; Asai et al., 2024; Jiang et al., 2023). Source attribution can show that a path came from the intended document while leaving the route-time version judgment unresolved.

Proposition 1 isolates exactly this missing information, and VERSIONSEAL carries it through cache validation, merge, and admissibility reconstruction.

Appendix Table 7 locates that distinction at mechanism level: prior families expose useful pieces of identity, snapshots, gating, or answer outcomes, while the route/path binding is the additional object studied here.

The distinctive composition is the request-scoped link across these columns. Proposition 1 constructs two histories for which final freshness and provenance agree but the route-snapshot judgment differs. VERSIONSEAL preserves the distinguishing generation as the authority for every returned key-version, turning otherwise separate routing, storage, and citation observations into one enforceable decision.

3 THE ROUTE–EVIDENCE RACE

We consider N sites that maintain asynchronously updated evidence graphs. A query arrives at coordinator time t_0; the coordinator reads routing state at gate-snapshot time t_g, commits the selected set at t_r, and receives evidence at t_s, with $t_0 < t_g \leq t_r < t_s$. Eligibility and relevance scoring consume the gate snapshot, and route commit seals that already-consumed state. Claim-key ownership is single-writer within an epoch, including the bounded migration experiment in Appendix G. Hybrid logical clocks totally order versions from one owner. Physical-clock error is bounded by ϵ_{clk}, which is charged against every declared freshness budget.

Definition 1 (Evidence claim)**.** *An evidence claim is represented by the tuple $c = (k, v, \tau, s)$, whose fields are the claim key k, the version identifier $v \in \mathbb{N}$, the commit timestamp τ, and the source identifier s. Within a fixed key and writer epoch, higher versions supersede lower ones.*

Site i publishes a monotonic watermark $W_i(t)$ and immutable index-generation identifier $G_i(t)$ in the same metadata transaction that makes all commits through $W_i(t)$ visible in its path index and routing sketch. W_i is the commit time of the newest update represented by G_i, not a liveness heartbeat; it remains fixed when no new write becomes visible. Let $V_i(k, G)$ denote the version of key k in generation G. For intent y with budget $\Delta_{\max}(y)$, site i is eligible at the gate snapshot exactly when

$$t_g - W_i(t_g) + \epsilon_{\text{clk}} \leq \Delta_{\max}(y). \tag{1}$$

After scoring only eligible sites, the coordinator commits the route and seals

$$e_q = \big(t_r, t_g, S_q, \{(W_i(t_g), G_i(t_g))\}_{i \in S_q}\big). \tag{2}$$

The digest-covered record also includes query ID, intent, budget, writer epochs, and coordinator term. Advancing a site after t_g does not change the generations authorized by this route.

Definition 2 (Cited-version violation)**.** *For a path returned by $i \in S_q$, the authorized version of key k is $v^\star = V_i(k, G_i(t_g))$. A cited pair (k, v_p) violates the route snapshot when $v_p \neq v^\star$. The mismatch is* superseded *when $v_p < v^\star$,* post-snapshot *when $v_p > v^\star$, and* unresolved *when the key is missing or tombstoned in the named generation.*

The experiments retain the established label stale-citation rate (SCR) for the combined exact-snapshot mismatch rate: violating returned paths divided by all returned paths. Thus SCR is an authorization metric, not a claim that every mismatch is chronologically old.

A path-cache entry stores a path signature, writer epoch, constituent (k, v) pairs, source IDs, score, and digest. Cache lookup is independent of e_q: an entry materialized under an earlier or later generation can match the same path signature. The generation in e_q supplies the authoritative map $V_i(\cdot, G_i(t_g))$; validation compares the cache payload with that map rather than assuming that a cache hit was materialized from it. This explicit join is where the two histories in Proposition 1 diverge.

Definition 3 (Route-admissible path)**.** *A served path is route-admissible when Eq. 1 held for its site, every cited key-version equals its $v^\star$, and the path digest is bound to e_q.*

Strict operation checks this predicate before service and fails closed on a mismatch. Balanced operation starts the same exact-generation lookup concurrently with cache service; a cache path that reaches the merge deadline while its lookup is outstanding may enter the response, and its later disposition remains recorded. Strict VERSIONSEAL instead completes the join before service. Proposition 2 states the resulting conditional served-path property under atomic publication, authentic tokens, and exact-version validation; Section 5 measures the enforcement–availability frontier between the two dispositions.

Metrics and dispositions. A returned path is any path merged into a generator context; a strict-served path has passed pre-service validation. The strict zero-event statistic and its upper bound use strict-served paths, whereas Balanced SCR uses all returned paths; neither denominator is query-level. UVL is commit-to-query-visibility time per update. Retrieval latency spans t_0 through merged paths and excludes generation, and WAN/query counts retrieval-layer bytes. Availability is served queries divided by issued queries. A strict mismatch triggers one reseal-and-retry, followed by abstention if no eligible site responds before the 800 ms per-site deadline; multi-second query tails can include parallel site waits, the one retry, coordinator work, and merge. Answer rates use all adjudicated questions as denominators.

Proposition 1 (Post hoc indistinguishability)**.** *Final site frontiers, selected-site identities, cited source identifiers, and post hoc provenance do not determine whether a returned key-version matched the route snapshot.*

Construction. Consider histories H_1 and H_2 that both end with versions $v_1 < v_2$ visible and both return a path citing v_1 from the same site and source. In H_1, v_2 is visible at t_g; in H_2, the same update becomes visible immediately after t_g. The final observations coincide, but Definition 2 rejects only H_1. The generation in e_q distinguishes the histories. Appendix A.1 gives the complete construction.

This separation identifies the state that an empirical control must preserve. FreshSketch supplies a stronger test than ordinary selective routing because it exports the route-visible watermark and returned version IDs as separate observations. VERSIONSEAL additionally names an authoritative generation in the route and joins each returned cache payload to it. Section 5.2 estimates the complete-protocol difference, and Table 6 resolves how publication, gating, identity, and record fields affect violation rate or admissibility reconstruction.

Executable decision sequence. The coordinator (1) reads a linearizable set of (W_i, G_i) pairs at t_g, applies Eq. 1, and scores only eligible sites; (2) it commits S_q at t_r by sealing those pairs in e_q; (3) each site retrieves version-addressed cache candidates and resolves their keys in the named generation; (4) Strict waits for every comparison before returning, whereas Balanced may return a deadline-ready candidate with the comparison outstanding; (5) the coordinator verifies token and path digests, merges by key-version, retries one rejected token, and otherwise serves or abstains; and (6) admissibility reconstruction recomputes site eligibility and each returned-path judgment from complete sealed records. Section 4 implements this sequence.

4 VERSIONSEAL

VERSIONSEAL implements route-bound evidence retrieval as four ordered state transitions. Publication creates a named authority; routing seals that authority; cache service returns an independently materialized, version-addressed path; and validation joins the payload back to the sealed generation before service or as a recorded asynchronous disposition. Figure 1 traces this authority from publication to generator context.

Version-addressed paths. An ingestion record (k, v, τ, s, ω) carries key, version, commit time, source, and writer epoch. Sites extract paths of depth at most three and cache path digests together with every constituent key-version. The cache is indexed by path signature and writer epoch, not by route token or index generation, so publication does not synchronously rematerialize every reusable path. The emitted identity names the object to be authorized; the named generation supplies the comparison target. Removing version IDs leaves source-level provenance intact but makes two revisions of the same source observationally identical.

Atomic visibility publication. Updates append to an ingestion log, compact into a query-visible RocksDB segment, update the routing sketch, and publish (W_i, G_i) in one transaction boundary. A crash before publication exposes neither the new watermark nor a partially advanced generation; recovery replays the unpublished batch. Generation G_i names the version map $V_i(\cdot, G_i)$ and remains retained for in-flight tokens. Decoupling the sketch, segment manifest, and watermark would let routing consume a state that path service cannot reproduce.

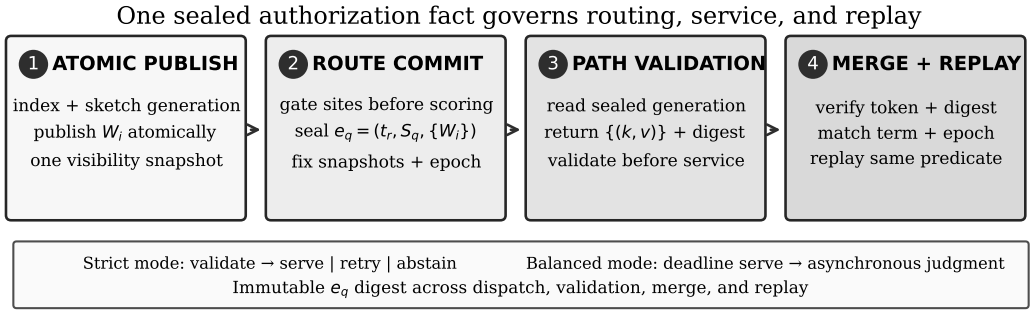


Figure 1: VERSIONSEAL's route-to-evidence dataflow; the diagram writes W_i as shorthand for the paired (W_i, G_i) gate snapshot and omits t_g for space. Atomic publication emits this pair, pre-score gating consumes it at t_g, and route commit seals the selected pairs at t_r. Cache paths carry exact versions; the token supplies the generation against which they are judged before Strict service or after asynchronous Balanced merge.

Eligibility before relevance. At t_g, the coordinator obtains (W_i, G_i) pairs from its three-replica Raft registry, removes sites that fail Eq. 1, and scores the remaining sketches. It then commits the selected set at t_r by sealing the exact pairs consumed by gating and scoring. This order makes freshness policy authoritative over relevance and preserves the gate-before-commit timeline. If no site qualifies, the coordinator refreshes the registry once and abstains after the site deadline. Writer-epoch and coordinator-term fields reject ownership and failover responses from outside the sealed route.

Path validation and replay. For each cache candidate, the site returns $(e_q, \text{site}, G_i, \{(k, v)\}, \text{pathDigest})$ and resolves the constituent keys in G_i. Strict mode waits for every equality $v = V_i(k, G_i)$ before releasing the candidate; mismatch, timeout, or term change rejects it. Balanced mode starts the same lookup concurrently but allows a cache hit to reach the merge deadline first, in which case the path may enter the response and the eventual pass/fail disposition is appended to its record. The merger verifies token and path digests, deduplicates at the key-version level, and selects the top B paths. Admissibility reconstruction reapplies the exact-generation predicate wherever the token and returned-path record are complete.

The stages make two outcomes separately testable. Atomic publication and pre-score gating determine which version state may compete, while version identity and token fields determine whether a returned path can be authorized and later reconstructed. Table 6 removes these observables individually: publication and gating change oracle SCR, whereas identity and record fields primarily change admissibility-reconstruction completeness.

The implementation uses RocksDB 8.6.7 for local segments, hybrid logical clocks (Lamport, 1978; Kulkarni et al., 2014), Bloom/min-hash routing summaries (Bloom, 1970; Broder, 1997), and Raft (Ongaro & Ousterhout, 2014). spaCy 3.7 plus domain patterns identifies entities, and a RoBERTa-base classifier extracts relations. These components instantiate the new cross-layer contract while keeping ranking and generation unchanged; Appendix B specifies transaction, cache, timeout, checkpoint, and reconstruction details.

Appendix Table 8 maps each protected decision to its assumption, failure disposition, and empirical stress, making the correspondence between protocol claims and evaluation evidence explicit.

5 EVALUATION

5.1 EXPERIMENTAL CONTRACT AND QUESTIONS

The evaluation follows four questions along the authorization trace: does the complete protocol reduce route-snapshot violations relative to freshness-aware selectivity (Q1); what availability does pre-service enforcement exchange for exact validation (Q2); do the retrieved contexts change temporal answers under a fixed generator (Q3); and which protocol stages control violations versus later

Table 1: FreshSketch–VERSIONSEAL protocol crosswalk. Shared rows hold retrieval capacity and ranking fixed; the final rows define the complete authorization-protocol contrast estimated in Table 2 and decomposed by Table 6.

Stage	FreshSketch	VERSIONSEAL
Candidates	Same sketch, budgets, pre-score gate, ranker, and selective fanout	Same
Paths and merge	Same version IDs, path cache, top-$B = 50$, scorer, and deduplication	Same
Timeout/tuning	Same deadline, one retry, validation split, and objective	Same
Cache lookup	Path signature and writer epoch; cache payload independent of route state	Same lookup and payload
Route authority	Watermark and cache generation observed independently	Atomic (W_i, G_i) publication; G_i sealed in route token
Path disposition	Version ID logged without request-scoped generation lookup	Cache payload joined to $V_i(\cdot, G_i)$ before service in Strict and concurrently in Balanced

admissibility reconstruction (Q4)? Unless a row explicitly says Strict, every VERSIONSEAL result uses Balanced validation.

The primary deployment spans US-East, US-West, and EU-Central, with four nodes per region. Every node has 32 vCPUs, 128 GB RAM, NVMe storage, and an NVIDIA A10 24 GB GPU. Ubuntu 22.04, CUDA 12.4, RocksDB 8.6.7, gRPC 1.58, Rust 1.78, and Go 1.22 are fixed across methods. Emulated inter-region mean RTTs are 62/78/142 ms, with 0.08% packet loss and a 10 Gbps site bottleneck.

The governed corpus contains 2.4M FDA, clinical-guideline, CVE/vendor, and de-identified incident documents, producing 18.7M claims over 4.2M entity pairs. Each six-hour open-loop replay issues 3,000 queries and approximately 43,200 updates, an average of 500 queries/hour and 120 updates/minute. Update arrivals follow a Poisson process whose base rate is normalized so that two-minute $5\times$ bursts every 30 minutes preserve that run total. Safety/dosage/general intents comprise 40/30/30%. Three paired update-stream seeds are hashes of the frozen query split and seed index. All methods receive identical schedules, extraction outputs, top-$B = 50$ context budgets, merge rules, questions, generator, prompt, and decoding; retrieved contexts are method-specific. Retrieval latency stops before the fixed Llama-3.1-8B-Instruct generator.

TTL-AnswerCache reuses generated answers until a tuned TTL expires; Broadcast-Fed queries every site; Partial-Replicate searches a fixed replicated subset; and SmartFed selects sites from compressed routing sketches without a query-pinned version snapshot. FreshSketch is an author-constructed same-stack control designed to isolate route-bound authorization. It uses the common routing sketch, query-visible watermark, version-addressed path cache, freshness budgets, pre-score gate, ranker, selective fanout, 800 ms site timeout, top-B merger, and one retry. Its cache is queried by the same path signature and returns the same key-version and source fields as VERSIONSEAL. The controlled difference is the complete authorization protocol: FreshSketch observes its watermark and cache state independently, whereas VERSIONSEAL atomically publishes a named generation, seals that generation in the request, and assigns each cache path an exact-generation disposition. Both receive the same held-out tuning objective over availability, SCR, p99 latency, and WAN cost. Appendix B gives the common configuration and tuning rules.

5.2 Q1: COMPLETE PROTOCOL VERSUS FRESHNESS-AWARE SELECTIVITY

FreshSketch closes much of the gap from ordinary selective routing, making it the relevant observation-only control. The complete VERSIONSEAL protocol reduces SCR from 2.12% to 1.18%; the paired difference is 0.94 pp [0.78,1.09] under a 30-minute-block bootstrap. The same comparison changes p99 latency from 4.84 to 4.62 s and WAN/query from 146 to 133 KB, so the lower violation rate does not require wider or slower retrieval in these runs. FreshSketch records 9540 violations among

Table 2: Primary 12-site comparison over three paired six-hour replay seeds. Latency, all-intent p99 UVL, and WAN/query are medians over seeds; SCR pools violating returned paths over all returned paths from the same runs. VERSIONSEAL is Balanced. Latency excludes generation, and UVL pools updates from the full 40/30/30 intent mix. Lower is better.

Method	p95 lat. (s)	p99 lat. (s)	All-intent p99 UVL (s)	SCR (paths, %)	WAN/Q (KB)
TTL-AnswerCache	1.92	3.84	291.4	7.62	38
Broadcast-Fed	3.41	7.23	19.6	2.86	412
Partial-Replicate	2.58	5.14	71.4	4.08	96
SmartFed	2.74	5.62	24.8	3.41	186
FreshSketch	2.39	4.84	10.9	2.12	146
VERSIONSEAL-Balanced	2.31	4.62	8.14	1.18	133

Table 3: Enforcement cost on the same 12-site replay. Balanced SCR uses all returned paths; the strict count and upper bound use paths that pass pre-service validation. Availability uses all issued queries.

Mode	Stale / evaluated	SCR (%)	Avail. (%)	p99 (s) / WAN (KB)
Balanced	–	1.18	99.2	4.62 / 133
Strict	0 / 443700	$\leq$0.00068	98.6	4.97 / 141

450000 returned paths, with a seed-paired 95% SCR interval of [1.96,2.31]. The 1.64-point safety-intent decrease (2.58% to 0.94%) is larger than the aggregate difference. Table 1 defines this as a complete-protocol estimand—atomic publication, route sealing, and path disposition act as one ordered contract—while Table 6 resolves the measured roles of publication, gating, identity, and record completeness.

5.3 Q2: PRE-SERVICE VALIDATION AND AVAILABILITY

Under the assumptions of Proposition 2, every strict-served path has passed the route-snapshot predicate. The online count records 0 violations among 443700 such paths; the one-sided 95% path-level upper bound is 0.00068%. Strict operation provides 98.6% availability at a 4.97 s p99, quantifying the cost of reseal/retry and abstention. Balanced operation raises availability to 99.2% and permits route/path admissibility reconstruction for 99.6% of emitted records, while retaining a measurable 1.18% mismatch rate.

5.4 Q3: TEMPORAL QA CONSEQUENCE

Annotators receive the question time, the answer, and all cited key-versions. *Temporally correct* requires every requested answer field to match the source state effective at the question time; partial or mixed-time answers fail. *Admissible citation* requires every cited key-version in the answer to satisfy its recorded route snapshot, so one mismatching citation fails the label. *Outdated/unsafe* marks an answer that asserts a superseded value or recommends an action conflicting with the effective dosage, safety, or vulnerability guidance. These are overlapping binary labels rather than a partition. Abstentions count as incorrect and without an admissible citation, but not as outdated/unsafe; disagreements are resolved by adjudication before aggregation.

Cluster bootstrap over paired update bursts gives a VERSIONSEAL-Balanced-minus-FreshSketch correctness effect of +4.6 pp [1.9,7.2] and an outdated/unsafe effect of -3.9 pp [-6.1,-1.8]. The paired burst-level point estimate is -3.9 pp; the pooled displayed counts give -3.8 pp after rounding. Holding model, decoding, questions, and context capacity fixed makes this the downstream effect of the complete retrieval protocol; method-specific site choice, ranking, and versions remain jointly represented in the retrieved context.

Table 4: Held-out temporal QA on the same 1200 entity- and time-disjoint questions for every method, equally split across safety, dosage, and general intents. Questions, generator, prompt, greedy decoding, context budget, and abstention rule are fixed; retrieved contexts differ by method. Columns are separately adjudicated binary outcomes over all questions. Agreement over the three-label decisions before adjudication is $\kappa = 0.84$, 95% CI [0.81,0.87].

Method	Temporally correct	Admissible citation	Outdated / unsafe
TTL-AnswerCache	824/1200 (68.7%)	953/1200 (79.4%)	227/1200 (18.9%)
Broadcast-Fed	898/1200 (74.8%)	1063/1200 (88.6%)	157/1200 (13.1%)
SmartFed	877/1200 (73.1%)	1049/1200 (87.4%)	178/1200 (14.8%)
FreshSketch	943/1200 (78.6%)	1110/1200 (92.5%)	116/1200 (9.7%)
VERSIONSEAL-Balanced	998/1200 (83.2%)	1153/1200 (96.1%)	70/1200 (5.8%)

Table 5: Five-seed, 14-day replay totaling 70000 queries and 6048000 Wikidata/Wikipedia revision updates over 16 logical sites. VERSIONSEAL is Balanced. SCR is path-level with seed-paired 95% intervals; UVL and retrieval latency are all-intent p99 seconds, and WAN is KB/query.

Method	SCR (%)	UVL	p99 lat.	WAN/Q
Broadcast-Fed	2.31 [2.13,2.49]	15.7	7.01	378
SmartFed	2.86 [2.63,3.11]	20.6	5.37	167
FreshSketch	1.72 [1.57,1.89]	9.4	4.71	142
VERSIONSEAL-Balanced	0.91 [0.82,1.01]	6.8	4.49	131

5.5 REVISION-DRIVEN PUBLIC REPLAY

Revision-driven updates preserve the main ordering: VERSIONSEAL-Balanced reduces SCR from 1.72 [1.57,1.89]% to 0.91 [0.82,1.01]% relative to FreshSketch, with lower UVL, retrieval tail, and traffic in the same replay.

5.6 Q4: WHICH STAGES CONTROL VIOLATIONS AND RECONSTRUCTION?

The ablation rejects a monolithic explanation. Removing version IDs or record fields barely changes oracle SCR but destroys admissibility reconstruction. By contrast, removing the watermark gate raises SCR to 4.72%, moving the gate after scoring raises it to 3.06%, and breaking atomic publication raises it to 2.47%. The measured 1.18% therefore depends on the ordered composition of publication and eligibility; identity and token metadata make each returned-path judgment exact and reconstructible. The budget sweep in Figure 3 separately traces policy: at $\Delta_{\max} = 5$ s, SCR is 0.72% and availability 96.2%; at the default 10 s, they are 1.18% and 99.2%.

5.7 DURATION, LOAD, FAILURES, AND MIGRATION

Ten paired 24-hour replays preserve the central ordering: median SCR (%, 95% interval) is 2.20 [2.02,2.39] for FreshSketch and 1.25 [1.14,1.37] for VERSIONSEAL-Balanced, a paired ratio of 0.57 [0.51,0.64]; Table 12 reports the four long-duration methods and their tails. At the directly measured 300-query/s load point, VERSIONSEAL-Balanced has a 4.88 s p99 versus 5.11 s for FreshSketch (Table 13).

Failure metrics come from a separate 256-client closed-loop harness, not the open-loop primary replay. During a 60 s regional outage, VERSIONSEAL-Balanced sustains 418 queries/s versus 312 for Broadcast-Fed and adds 0.7 s rather than 3.5 s to p99 latency. The Balanced coordinator-failover slice counts only safety queries during election: election is 1.18 s median (2.4 s p99), and asynchronous validation leaves 7 of 892 queries with at least one stale citation. Strict applies the same old-term check before service and rejects these responses, converting the event into retry or abstention. Appendix H reports the two windows separately. With 5% of keys migrating ownership during a burst, VERSIONSEAL-Balanced holds SCR to 1.46% versus 3.18% for FreshSketch (Table 14).

Table 6: Leave-one-observable-out attribution on the primary replay. Oracle SCR adjudicates paths from complete version records; reconstruction is the fraction of returned route/path admissibility judgments recoverable from emitted records. Latency is p99 seconds and WAN is KB/query.

Variant	SCR	Recon. (%)	p99	WAN	Failure disposition
Full Balanced	1.18	99.6	4.62	133	Detect
No version IDs	1.19	0.0	4.55	129	Unverifiable
No watermark/gate	4.72	95.4	5.08	171	Serve stale
Gate after scoring	3.06	97.8	4.73	148	Late reject
No record metadata	1.21	18.3	4.58	127	Unlocalized
Non-atomic publish	2.47	92.1	4.66	136	Reject race

6 Discussion

The results separate three decisions that freshness-oriented retrieval often collapses: whether a site may compete, whether a returned path matches the state that admitted it, and whether that judgment can be reconstructed later. Publication and pre-score gating control the first two decisions' shared version state, while exact key-version identity and sealed record fields make the path-level judgment executable. Strict and Balanced use the same predicate but differ in when its disposition can affect service; the measured frontier is therefore an enforcement-scheduling choice rather than a change in authorization semantics. The temporal-QA effect shows why this distinction matters beyond observability: route-inconsistent contexts alter generated answers even when model and context capacity are fixed.

The evaluated contract covers authenticated route records, bounded clock error, retained named generations, and one writer per key within each sealed epoch; migration and coordinator-failover stresses exercise changes in ownership epoch and authority term across three emulated-WAN regions. Simultaneously writable replicas would additionally require the token to name the conflict-resolution state that defines $V_i(k, G)$ before the same admissibility predicate can be applied. Within the evaluated regime, the primary and public revision replays, ten paired 24-hour runs, migration stress, and temporal QA consistently establish the value of route-bound exact-version authorization and quantify its enforcement–availability frontier.

7 Conclusion

The same final citation can encode opposite authorization judgments because its validity depends on the generation that routing consumed. VersionSeal preserves that otherwise-lost fact: atomic publication names the visibility state, route commit seals it, cache paths expose exact key-versions, and one predicate governs Strict service and every later admissibility reconstruction. In the evaluated three-region single-writer regime, this complete protocol reduces SCR relative to FreshSketch by 0.94 pp [0.78,1.09]; Strict records 0 violations among 443700 served paths, while Balanced provides 99.2% availability and improves temporal-answer correctness by +4.6 pp [1.9,7.2]. The public revision replay and ten paired 24-hour runs preserve the ordering. Mutable retrieval systems should therefore treat route commit as an evidence-authorization boundary, not merely a source-selection event.

Ethics Statement

The evaluation uses public regulatory, clinical-guideline, and vulnerability records together with de-identified incident documents, and it does not make decisions about individuals. Human evaluation was limited to blinded annotation of system outputs; only aggregate judgments are reported. Data handling and annotation followed the applicable access and institutional requirements.

AI USE STATEMENT

Generative AI tools were used for language editing, presentation organization, reference formatting, and submission-readiness checks. The authors reviewed all AI-assisted work and take responsibility for the final content of this manuscript.

REFERENCES

Tyler Akidau, Alex Balikov, Kaya Bekiroglu, Slava Chernyak, Josh Haberman, Reuven Lax, Sam McVeety, Daniel Mills, Paul Nordstrom, and Sam Whittle. Millwheel: Fault-tolerant stream processing at internet scale. *Proceedings of the VLDB Endowment*, 6(11):1033–1044, 2013. doi: 10.14778/2536222.2536229.

Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle. The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. *Proceedings of the VLDB Endowment*, 8(12):1792–1803, 2015. doi: 10.14778/2824032.2824076.

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

Peter Bailis, Shivaram Venkataraman, Michael J. Franklin, Joseph M. Hellerstein, and Ion Stoica. Probabilistically bounded staleness for practical partial quorums. *Proceedings of the VLDB Endowment*, 5(8):776–787, 2012. doi: 10.14778/2212351.2212359.

Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970. doi: 10.1145/362686.362692.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pp. 2206–2240. PMLR, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Andrei Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences (SEQUENCES)*, 1997. doi: 10.1109/SEQUEN.1997.666900.

James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 21–28. Association for Computing Machinery, 1995. doi: 10.1145/215206.215328.

Paris Carbone, Stephan Ewen, Seif Haridi, Asterios Katsifodimos, Volker Markl, and Kostas Tzoumas. Apache Flink: Stream and batch processing in a single engine. *IEEE Data Engineering Bulletin*, 38 (4):28–38, 2015. URL https://sites.computer.org/debull/A15dec/p28.pdf.

James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. Spanner: Google's globally-distributed database. In *10th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 261–264. USENIX Association, 2012. URL https://www.usenix.org/conference/osdi12/technical-sessions/presentation/corbett.

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: Amazon's highly available key-value store. In *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP)*, pp. 205–220. Association for Computing Machinery, 2007. doi: 10.1145/1294261.1294281.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8472–8494. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.496. URL https://aclanthology.org/2024.findings-emnlp.496/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Rachid Guerraoui, Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, and Martijn de Vos. Efficient federated search for retrieval-augmented generation. In *Proceedings of the 5th Workshop on Machine Learning and Systems (EuroMLSys)*. Association for Computing Machinery, 2025. doi: 10.1145/3721146.3721942.

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

Daniel Huwiler, Kurt Stockinger, and Jonathan Fürst. VersionRAG: Version-aware retrieval-augmented generation for evolving documents. *arXiv preprint arXiv:2510.08109*, 2025. doi: 10.48550/arXiv.2510.08109. URL https://arxiv.org/abs/2510.08109.

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.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Anagha Kulkarni and Jamie Callan. Selective search: Efficient and effective search of large textual collections. *ACM Transactions on Information Systems (TOIS)*, 33(4):17:1–17:33, 2015. doi: 10.1145/2738035.

Sandeep S. Kulkarni, Murat Demirbas, Deepak Madappa, Bharadwaj Avva, and Marcelo Leone. Logical physical clocks. In *International Conference on Principles of Distributed Systems (OPODIS)*, volume 8878 of *Lecture Notes in Computer Science*, pp. 17–32, 2014. doi: 10.1007/978-3-319-14472-6_2.

Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565, 1978. doi: 10.1145/359545.359563. URL https://doi.org/10.1145/359545.359563.

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

Wyatt Lloyd, Michael J. Freedman, Michael Kaminsky, and David G. Andersen. Don't settle for eventual: Scalable causal consistency for wide-area storage with COPS. In *Proceedings of the 23rd ACM Symposium on Operating Systems Principles (SOSP)*, pp. 401–416. Association for Computing Machinery, 2011. doi: 10.1145/2043556.2043593.

Diego Ongaro and John Ousterhout. In search of an understandable consensus algorithm. In *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, pp. 305–319. USENIX Association, 2014. URL https://www.usenix.org/conference/atc14/technical-sessions/presentation/ongaro.

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

Luo Si and James P. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 298–305. Association for Computing Machinery, 2003. doi: 10.1145/860435.860490.

Alessandro Stolfo. Groundedness in retrieval-augmented long-form generation: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1537–1552. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.100. URL https://aclanthology.org/2024.findings-naacl.100/.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, and Yong Liao. EvoWiki: Evaluating LLMs on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 948–964. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.47. URL https://aclanthology.org/2025.acl-long.47/.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.813.

Shuai Wang, Ekaterina Khramtsova, Shengyao Zhuang, and Guido Zuccon. FeB4RAG: Evaluating federated search in the context of retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 763–773. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657853.

Feifan Wu, Lingyuan Liu, Wentao He, Ziqi Liu, Zhiqiang Zhang, Haofen Wang, and Meng Wang. Time-sensitive retrieval-augmented generation for question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2544–2553. Association for Computing Machinery, 2024. doi: 10.1145/3627673.3679800.

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

A Proofs

A.1 Post hoc indistinguishability

Fix one site i, key k, source s, versions $v_1 < v_2$, and gate-snapshot time t_g. In history H_1, the transaction publishing v_2, watermark $W_i = \tau_2$, and generation $G_i^{(2)}$ commits before t_g; an

asynchronous path cache nevertheless returns (k, v_1, s). In H_2, the same publication commits immediately after t_g, so the route snapshot names $G_i^{(1)}$ and the cache correctly returns (k, v_1, s). Allow both executions to finish before observation. Their final graph, final watermark, selected site, returned source, and post hoc provenance tuple are identical. Definition 2 rejects H_1 because $V_i(k, G_i^{(2)}) = v_2$ and accepts H_2 because $V_i(k, G_i^{(1)}) = v_1$. The route tokens contain different generation identifiers and therefore distinguish the executions, proving Proposition 1.

A.2 Strict served-path compliance

Proposition 2 (Conditional strict authorization compliance)**.** *Assume single-writer-per-epoch versions, atomic index/sketch/watermark publication, an authentic route token, and strict site-local validation against the token snapshot. Every path served by strict mode satisfies Definition 3.*

Proof. Atomic publication makes $V_i(k, G_i(t_g))$ a well-defined query-visible version for every selected site and key. Eq. 1 holds because the coordinator forms S_q only after eligibility filtering at t_g. Before service, strict validation compares every returned (k, v_p) with $V_i(k, G_i(t_g))$ and rejects the path if they differ. The site also binds the path digest to the authentic token, and the coordinator rejects term, epoch, generation, or digest mismatch. Every served path therefore satisfies site eligibility, exact version equality, and token binding, which are precisely the clauses of Definition 3. Rejected or timed-out queries affect availability, not the served-path conclusion. □

B Implementation and Evaluation Protocol

Table 7: Mechanism-level comparison. ID: exact cited-version identity; Snap.: named routing snapshot; Gate: pre-score eligibility; Bind: route/path binding; Mut.: mutable-corpus evaluation; Ans.: temporal answer outcomes. A circle marks an unbound or partial observable. FreshSketch is our same-stack control.

Family	ID	Snap.	Gate	Bind	Mut.	Ans.
Dynamic / version-aware RAG	○	○	○	–	✓	✓
Federated RAG routing	–	–	–	–	○	✓
Watermark / bounded-staleness systems	○	✓	–	–	✓	–
Citation / provenance tracing	○	–	–	–	○	✓
FreshSketch	✓	✓	✓	–	✓	✓
VersionSeal	✓	✓	✓	✓	✓	✓

Each update batch is appended to the ingestion log, built into an immutable RocksDB segment, and applied to a shadow routing sketch. A single metadata transaction installs the segment manifest, swaps the sketch generation, and advances W_i; recovery observes either all three new pointers or the preceding generation. Path-cache entries are keyed by path signature and writer epoch and retain the constituent versions from their materialization event. `RetrievePaths` receives the sealed token, obtains cache candidates, and resolves their keys in the token's named generation. Strict mode waits for every version comparison before returning; Balanced starts the same lookup concurrently and may release a deadline-ready cache hit while the comparison is outstanding. A timeout causes the coordinator to refresh registry state and seal one new token. A second timeout, an empty eligible set, a changed writer epoch, a missing retained generation, or an invalid path digest produces abstention.

Freshness budgets were selected before the evaluation by service class: 10 s for safety, 30 s for dosage, and 60 s for general queries. Baselines receive the same validation split and tune against the same lexicographic objective: satisfy the availability target, then minimize SCR, p99 retrieval latency, and WAN/query. Evaluation entities, timestamps, and paraphrase templates are disjoint from tuning. The three primary seeds are SHA-256 hashes of the fixed split manifest concatenated with seed index; paired methods consume identical query and update schedules. FreshSketch uses the common sketch, ranker, cache keys and payloads, fanout floor, site deadline, one-retry policy, and merger. It reads a query-visible watermark for gating and returns version IDs, but its request contains no authoritative generation identifier and it performs no request-scoped lookup joining those IDs to the gate state.

Table 8: Claim-to-evidence map. Each row identifies a protocol decision, its assumption and checked property, and the experiment that exercises it. "Reject" includes one reseal-and-retry before abstention.

Protected decision / event	Assumption and checked property	Comparator or stress; disposition; evidence
Index visibility / delayed compaction	Bounded site clock; (W_i, G_i) names one visible state	FreshSketch and 24-hour bursts; gate or reject; Tables 2, 12
Routing state / stale sketch	Authenticated registry; freshness precedes scoring	No-watermark and post-score ablations; reseal; Table 6
Key-version / route–serve race	Named-generation lookup; cache path matches sealed snapshot	FreshSketch and held-out QA updates; validate or reject; Tables 3, 4
Route token / regional timeout	Stable selected set and coordinator term	Balanced closed-loop outage; retry or abstain; Appendix H
Coordinator term / failover	Linearizable token registry; old terms rejected	Balanced leader-kill slice; reject old term asynchronously; Section 5.7
Admissibility record / omitted field	Persisted token; judgment reconstructible	Leave-one-field-out reconstruction; mark incomplete; Table 6
Writer epoch / migration	One owner per token epoch; old epochs rejected	5% burst migration; reject old epoch; Table 14

Table 9: Route-token and returned-path schemas. Fields covered by the token digest define the authorization record and remain immutable across dispatch, site validation, merge, and admissibility reconstruction.

Record	Fields
Route token e_q	query ID; gate-snapshot and route-commit HLCs; intent; freshness budget; selected site IDs; site watermarks and immutable index-generation IDs; writer epochs; coordinator term; retry generation
Returned path	route-token digest; site ID; index-generation ID; ordered claim keys and versions; source IDs; path score; path digest; validation disposition
Reconstruction output	site eligibility; token-visible version per key; path admissibility; rejection reason; merge position

Statistical reporting. Primary SCR point estimates pool returned paths from each set of three paired runs. The paired difference resamples 30-minute update windows within each shared schedule and preserves method pairing; paths are never treated as independent runs. The temporal-QA intervals analogously resample paired questions by update burst. Long-horizon and public-replay SCR intervals use their stated independent paired runs or seeds. Path-level zero-event bounds describe the evaluated served set and are not cluster-adjusted population estimates.

The extraction stack uses spaCy 3.7 NER with medical and CVE patterns and the `FacebookAI/roberta-base` checkpoint fine-tuned for relation classification. Retrieval uses the same embeddings, evidence-path candidates, and merge scorer for every method. Answer evaluation uses `meta-llama/Meta-Llama-3.1-8B-Instruct` with greedy decoding, a fixed prompt, and the same top-50 evidence budget.

Temporal-QA adjudication. The two blinded annotators independently label three binary properties using the question timestamp and the effective source revision. Temporal correctness requires all requested fields to match that revision; citation admissibility requires every cited (k, v) to match its route snapshot; and outdated/unsafe marks either a superseded factual value or action-bearing text that conflicts with the effective safety, dosage, or vulnerability guidance. Multiple citations use an all-citations rule, and partial answers use an all-requested-fields rule. Abstentions receive zero for correctness and citation admissibility and are not counted as outdated/unsafe. The reported agreement summarizes the three-label decisions before an adjudicator resolves disagreements.

Table 10: Default VERSIONSEAL configuration used unless an experiment varies the named parameter.

Component	Parameter	Value
Site index	LSM memtable / compaction trigger	64 / 128 MB
	Max path depth / top-B	3 hops / 50
Sketch	Bloom filter / false-positive rate	1 MB / 0.01
	Heavy-hitter K / push interval	100 / 2 s
Coordinator	Raft replicas / follower lease	3 / 2 s
	Site timeout / exploration floor	800 ms / 2 sites
Freshness	Safety / dosage / general budget	10 / 30 / 60 s

C PRIMARY-SETTING BREAKDOWNS

Table 11: Path-level SCR (%) by intent on the 12-site, three-seed replay used in Table 2. Overall values use exact raw totals; displayed intent cells are rounded independently.

Method	Safety	Dosage	General	Overall
TTL-AnswerCache	9.28	7.14	5.88	7.62
Broadcast-Fed	3.12	2.79	2.61	2.86
Partial-Replicate	4.64	4.01	3.43	4.08
SmartFed	3.92	3.33	2.82	3.41
FreshSketch	2.58	2.07	1.55	2.12
VERSIONSEAL-Balanced	0.94	1.15	1.52	1.18

The intent breakdown exposes where route-time identity matters most: VERSIONSEAL-Balanced lowers safety SCR from 2.58% for FreshSketch to 0.94%, while the general-intent rates are nearly equal. This ordering matches the tighter 10 s safety budget and motivates the safety-query UVL view in Figure 2.

Together, the CDF and sweep separate systems delay from authorization policy: VERSIONSEAL's retrieval tail remains competitive, while the freshness budget moves SCR and availability along an explicit operating frontier. The no-watermark intervention moves off that frontier by removing the state used for pre-score eligibility.

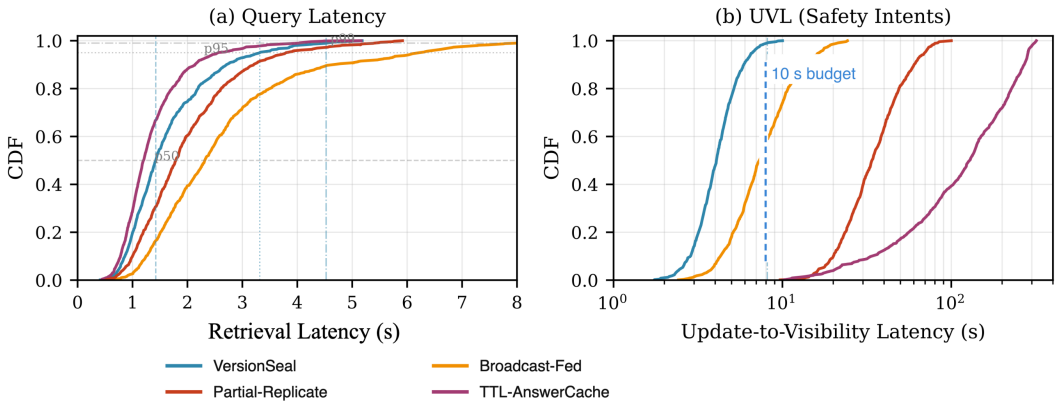


Figure 2: Empirical CDFs over the three paired six-hour runs: retrieval latency for all queries (left) and the safety-query UVL distribution under the 10 s budget (right). The latter population is distinct from the all-intent p99 UVL cells in Table 2.

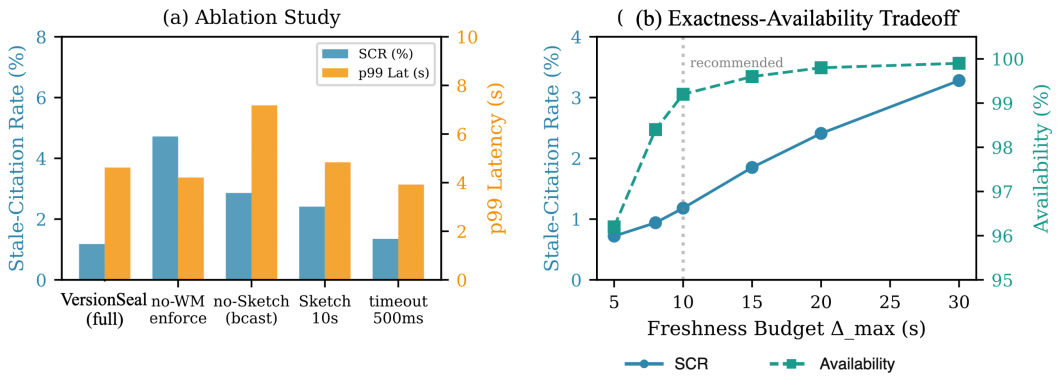


Figure 3: Primary-replay mechanism and budget sweeps. Removing watermark enforcement raises path-level SCR from 1.18% to 4.72%; changing $\Delta_{\max}$ traces the measured availability–freshness frontier around the 10 s safety operating point.

D PUBLIC TEMPORAL REPLAY

The public setting replays timestamped Wikidata statements and linked English Wikipedia revisions over a contiguous 14-day window. Claim versions are derived from revision IDs and commit timestamps; query templates are generated from changed relations and evaluated only on entities and timestamps disjoint from construction. Sixteen logical sites partition entities by stable hash. Five seeds preserve the revision stream and vary held-out query arrival times, producing 70000 queries and 6048000 update events in total. Every method uses the same path extractor, routing objective, context budget, and temporal path-adjudication rule as the primary experiment; all VERSIONSEAL results in this replay use Balanced validation.

E LONG-HORIZON ROBUSTNESS

Ten paired 24-hour runs contain 120000 queries and 1728000 updates in aggregate. As in the primary replay, the update process is normalized to the reported total before recurring bursts are applied. Methods share schedules within each pair. We block-bootstrap 30-minute burst windows and report medians with 95% intervals for SCR; UVL and latency are p99 over the aggregate replay.

Table 12: Long-horizon primary-setting robustness over ten paired 24-hour runs. VERSIONSEAL is Balanced. SCR is path-level percent with block-bootstrap 95% intervals; UVL and retrieval latency are p99 seconds.

Method	SCR (%)	p99 UVL	p99 latency
TTL-AnswerCache	7.79 [7.34,8.26]	294.8	3.91
SmartFed	3.56 [3.29,3.84]	25.5	5.71
FreshSketch	2.20 [2.02,2.39]	11.2	4.91
VERSIONSEAL-Balanced	1.25 [1.14,1.37]	8.36	4.70

The paired VERSIONSEAL/FreshSketch SCR ratio is 0.57 [0.51,0.64], preserving the main comparison across longer update tails.

F LOAD

The normal-load sweep uses offered rates of 100, 200, 300, 400, and 500 queries/s at a fixed update rate and query mix. Each directly measured point follows warm-up with five ten-minute plateaus; all four methods run on the identical 12-node stack. The common 300-query/s point supplies the method comparison reported in Table 13.

Table 13: Normal-load p99 retrieval latency at 300 queries/s on the 12-node stack. VERSIONSEAL is Balanced. Values are medians over five ten-minute post-warm-up plateaus with 95% plateau intervals.

Method	p99 at 300 QPS (s)
Broadcast-Fed	7.51 [7.32,7.74]
SmartFed	5.96 [5.80,6.14]
FreshSketch	5.11 [4.98,5.25]
VERSIONSEAL-Balanced	4.88 [4.76,5.03]

G OWNERSHIP MIGRATION

Migration preserves one writer per key and epoch. The coordinator seals writer epoch in e_q; old-epoch responses are rejected and retried after the ownership map advances. The migration-specific Last-writer route comparator follows the ownership map available at service time but carries neither a sealed writer epoch nor a route token.

Table 14: Ownership-migration stress during update bursts. VERSIONSEAL is Balanced. SCR is path-level percent at 0/1/5% migrating keys. Cost columns report the 5% condition; p99 overhead is relative to each method's no-migration point and metadata is bytes per returned path. Methods without route tokens cannot reject an obsolete epoch.

Method	SCR 0%	SCR 1%	SCR 5%	Reject (%)	p99 Δ (s)	Meta. (B)
Last-writer route	3.44	4.18	6.72	–	0.00	0
FreshSketch	2.14	2.37	3.18	–	0.06	16
VERSIONSEAL-Balanced	1.18	1.26	1.46	1.7	0.17	24

H BOUNDED FAILURES

The Balanced failure harness uses 256 closed-loop clients. During a 60 s regional outage, VERSIONSEAL-Balanced sustains 418 queries/s and adds 0.7 s to p99 retrieval latency, compared with 312 queries/s and 3.5 s for Broadcast-Fed. A separate coordinator-election window contains 892 safety queries; asynchronous validation leaves 7 with at least one stale citation, and election duration is 1.18 s median and 2.4 s p99. Strict checks the coordinator term before service and converts those old-term responses into retry or abstention.