MISLEADING RECENCY IN TEMPORAL RAG: INTENT-CONDITIONED EVIDENCE PACKETS UNDER FIXED RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

A query about Python 3.8 can retrieve both the requested tutorial and a newer Python 3.12 deprecation notice: publication chronology favors the notice, whereas semantic coherence with the requested version favors the tutorial. Correct handling is not a single recency correction. Wrong-Newer, Missing-Timestamp, Branch-Specific, and Partial-Rollback conflicts require different decisions about whether retrieved endpoints should be rejected, traversed, or retained together for a requested state, history, branch, or comparison. CHRONOPACKET makes that decision with typed claim relations and a bounded, intent-conditioned packet objective: supersession collapses revisions for state queries, while comparison and branch intents preserve the requested alternatives. To isolate packet commitment from retrieval, our 1,800-query protocol admits only examples whose shared top-20 pool contains a complete support set and identifies 240 misleading-recency test cases by timestamp-first incompatibility before any learned-selector output is observed. Diagnostic gains over full-pool reading are 5.9 points on Wrong-Newer, 4.6 on Missing-Timestamp, 6.3 on Branch-Specific, and 4.1 on Partial-Rollback queries. Across all misleading-recency cases, CHRONOPACKET reaches 66.4 $\pm$ 0.8% accuracy versus 65.1 $\pm$ 0.8% for Reader-Conditioned Selector (paired difference 1.3 points, 95% CI [0.4, 2.3]); the no-abstention control also lowers stale citations from 9.1% to 8.1% and unresolved contradictions from 10.7% to 7.0%, while answered-query median post-retrieval latency falls from 171 to 61 ms. Collapsing coexistence removes 8.7 Branch-Specific points and shuffling validity intervals removes 6.6 aggregate points; source and entity holdouts, independent transfer cases, faster incident updates, and BGE-M3 retrieval preserve the ordering. Under fixed answer-bearing access, these results identify intent-sensitive conflict topology as the decision unit for choosing which retrieved revisions should jointly govern a temporal RAG answer.

1 INTRODUCTION

Consider the citation trail for "What does `get_event_loop()` do in Python 3.8?" A retrieved Python 3.8 tutorial describes the requested behavior; a later Python 3.12 notice deprecates that behavior under a different version scope. Chronology ranks the notice first, but the tutorial is the governing evidence for the query. Supplying both passages to a reader does not resolve this disagreement: the out-of-scope revision can determine both the generated answer and the citation. The decision is therefore over a coherent set of retrieved claims, after answer-bearing access has already succeeded.

The required set operation changes with conflict topology. *Wrong-Newer* calls for rejecting a later false, premature, or transient claim. *Missing-Timestamp* requires recovering semantic validity when metadata order is absent or misleading. *Branch-Specific* requires preserving legitimate alternatives before selecting the branch named by the query, while *Partial-Rollback* requires following a directional chain to a restored state. A freshness rule cannot express rollback, a generic contradiction filter can erase endpoints needed for comparison, and unrestricted full-pool reading leaves scope resolution to generation. Current-state, historical, transition, comparison, and dispute intents can consequently

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.

map the same claim pair to different compatible evidence sets. We call the selected support set the *evidence commitment*.

Our evaluation holds evidence access constant and varies only this commitment rule. From 1,800 queries over API documentation, policy revisions, and entity-status histories, every admitted example contains a complete support set of at most four passages in the same top-20 pool. The 240-case misleading-recency (MR) test stratum is determined by timestamp-first failure, and source history plus query scope assigns one of the four topologies before learned selectors or readers are evaluated. Recency, NLI, passage-utility selectors, full-pool reading, and typed packet selection then share candidates, reader, prompt, and decoding. Answer accuracy, citation validity, unresolved conflict, coverage, and latency expose which policy turns identical access into a reliable answer.

CHRONOPACKET implements an intent-conditioned set decision before generation. It extracts scoped claims, predicts SUPPORT, SUPERSEDE, CONTRADICT, PARTIALCONFLICT, and COEXIST relations, infers one of five query intents, and approximately optimizes a four-passage objective over relevance, compatible edge patterns, and unresolved conflict. The central operation is relation reversal: a supersession edge retains one query-valid endpoint for a state query but both requested endpoints for a transition, while coexistence preserves branch alternatives before scope selects among them. The deployed greedy policy returns 2.8 passages on average and reaches the reader in 61 ms median post-retrieval time.

Wrong-Newer and Branch-Specific cases produce the largest topology-specific gains under identical access: relative to Long-Context Reader, CHRONOPACKET gains 5.9 and 6.3 points, while Missing-Timestamp and Partial-Rollback gain 4.6 and 4.1. Collapsing supersession damages Wrong-Newer and rollback decisions, whereas collapsing coexistence removes 8.7 Branch-Specific points. Across all MR cases, Reader-Conditioned Selector reaches $65.1 \pm 0.8\%$ and CHRONOPACKET $66.4 \pm 0.8\%$ (paired difference 1.3 points, 95% CI [0.4, 2.3]); the typed packets also reduce unresolved contradictions from 10.7% to 6.3% and stale citations from 9.1% to 7.8%. Against full-pool reading, MR accuracy rises by 5.1 points as stale citations fall by 5.4 points and answered-query latency by 223 ms. Source-family and entity holdouts, independent transfer construction, faster incident updates, and BGE-M3 retrieval preserve the ordering.

Contributions. The primary contribution is an access-controlled diagnostic that turns misleading recency into four auditable conflict topologies and measures the answer, citation, coverage, and cost produced by each commitment rule. The technical contribution is CHRONOPACKET's typed packet objective, whose intent-dependent supersession and coexistence semantics represent branch, rollback, and comparison decisions that passage-wise utility cannot. The empirical contribution is a matched-policy study with targeted relation interventions and independence checks that connects the observed reliability gains to those set operations rather than retrieval depth or packet-search capacity.

2 RELATED WORK

Temporal QA benchmarks such as SituatedQA, StreamingQA, RealTimeQA, TempLAMA, and FreshQA measure access to answers that change over time (Zhang & Choi, 2021; Liska et al., 2022; Kasai et al., 2023; Dhingra et al., 2022; Vu et al., 2024). T-GRAG, Temporal GraphRAG, and IA-RAG restructure retrieval with temporal graphs, query decomposition, or interval algebra (Li et al., 2025; Han et al., 2025; Wang et al., 2026b); MemStrata retires superseded facts before retrieval, and Re3 combines time-aware retrieval with listwise suppression of obsolete versions (Yadav, 2026; Cao et al., 2026). Re3's obsolete-version suppression targets current-state evidence, whereas our branch, comparison, and rollback cases require non-monotone retention decisions in which an older or alternative endpoint can remain necessary. More generally, methods that change access and evidence choice jointly cannot isolate which stage produced a temporal gain. Our fixed-access protocol retains competing snapshots in one gold-containing pool and measures only the commitment decision.

Rerankers and RankRAG score passage utility, while retrieval-conditioned readers consume multiple passages without an explicit temporal set decision (Nogueira & Cho, 2019; Izacard & Grave, 2021; Ram et al., 2023; Asai et al., 2024; Yu et al., 2024). Set selection additionally models collective coverage (Lee et al., 2025), and CalibRAG scores documents by predicted decision correctness (Jang et al., 2025). Our cross-encoder, reader-conditioned, and parameter-matched graph-free controls instantiate these alternatives inside the frozen pool. CHRONOPACKET's set criterion targets a separate

interaction: intent changes temporal compatibility. Directional supersession favors one endpoint for a state query, whereas branch and version comparisons reward retaining the requested alternatives. Typed-edge and intent interventions directly test this distinction under a common packet budget.

Counterfactual retrieval and robust-RAG studies demonstrate damage from relevant-looking context (Hong et al., 2024; Xiang et al., 2024; Zeng et al., 2025). FaithfulRAG, ASTUTE RAG, and MADAM-RAG model context–parameter or multi-source disagreement (Zhang et al., 2025; Wang et al., 2025a;b). ArbGraph performs pre-generation claim-graph arbitration for long-form factual consistency; ConflictRAG resolves generic document conflicts; and EvoTrustRAG attributes conflict to evolution, intervention, or uncertainty and adapts evidence presentation to the query (Niu et al., 2026; Wang et al., 2026a; Nie et al., 2026). EvoTrustRAG is closest in making conflict presentation query-dependent; our protocol instead fixes answer-bearing access, and our objective specifies the endpoint-retention pattern for each temporal relation and intent. ReliabilityRAG selects a consistent majority from a contradiction graph, while SeCon-RAG filters semantic and model conflicts in the corpus and retrieved context (Shen et al., 2025; Si et al., 2025). A claim pair that must collapse for one requested state may be necessary evidence for a comparison or dispute, and a branch or rollback cannot be resolved by generic credibility propagation alone. The MR protocol labels these topologies from source histories, while the method encodes their intent-dependent endpoint semantics. Calibration methods (Naeini et al., 2015; Guo et al., 2017) provide the selective decision when no packet clears the coherence threshold.

3 Evidence Access and Commitment

For a query q whose temporal intent is implicit or explicit, retrieval produces k passages $\mathcal{P} = \{p_1, \ldots, p_k\}$ drawn from one or more temporal snapshots. Let $\mathfrak{G}(q) \subseteq 2^{\mathcal{U}}$ be the family of query-compatible passage packets in corpus universe $\mathcal{U}$: each member contains a complete support set for the gold answer, every answer-bearing claim in the member is temporally and scope compatible, and unrelated passages are permitted. Evidence access succeeds iff

$$\exists\, G \in \mathfrak{G}(q) \text{ such that } G \subseteq \mathcal{P},$$

which requires a complete compatible packet rather than a nonempty overlap with one. A policy h makes a pre-generation evidence commitment $\mathcal{K}_h(q, \mathcal{P}) \subseteq \mathcal{P}$, the packet delivered to the common reader; the full-context policy commits all of $\mathcal{P}$. Commitment is query-compatible exactly when $\mathcal{K}_h(q, \mathcal{P}) \in \mathfrak{G}(q)$; answer and citation metrics separately evaluate the reader output and its cited support. The system may instead return an abstention symbol $\perp$ when no packet achieves sufficient coherence.

Queries fall into five temporal intent classes, $\mathcal{I} = \{$Current, AsOfTime, ChangeOverTime, VersionCompare, Disputed$\}$, which induce different coherence constraints. Branch is represented as a scope qualifier within these intents: a named-branch state query is Current or AsOfTime, whereas a branch comparison is VersionCompare. Each passage p_i is decomposed into claims $\mathcal{C}_i = \{c_1, \ldots, c_m\}$, where a claim $c_j = (s_j, r_j, o_j, \mathbf{q}_j, [t_j^{\text{start}}, t_j^{\text{end}}], \rho_j)$ stores the subject, relation, object, scope qualifiers, validity interval, and calibrated extraction confidence used as provenance confidence.

Let h_{time} denote a deterministic timestamp-first packet policy: passages with observed publication times are ordered newest first, undated passages retain their frozen retrieval order after dated passages, and equal timestamps are broken by frozen retrieval rank and then source order. Its commitment $\mathcal{K}_{h_{\text{time}}}(q, \mathcal{P})$ is the first B passages under this ordering. The misleading-recency (MR) stress stratum is the recency-conditioned access–commitment separation

$$\big[\exists G \in \mathfrak{G}(q) : G \subseteq \mathcal{P}\big] \quad \wedge \quad \mathcal{K}_{h_{\text{time}}}(q, \mathcal{P}) \notin \mathfrak{G}(q).$$

MR is therefore the conditional regime in which answer-bearing access and chronology disagree. Four structural indicators identify why recency fails. Wrong-Newer requires an incompatible claim published after a compatible claim; Missing-Timestamp requires publication order to be absent or inconsistent with claim validity; Branch-Specific requires qualifier-disjoint claims and a query naming one branch; Partial-Rollback requires a directional chain $c_a \prec c_b \prec c_c$ in which c_c restores the queried field of c_a after c_b. Raw indicators can overlap, so final labels apply the fixed precedence Partial-Rollback, Branch-Specific, Missing-Timestamp, then Wrong-Newer. Source histories, validity intervals, and query scope fix these labels before learned-selector and reader outputs are observed.

4 CHRONOPACKET: QUERY-CONDITIONED EVIDENCE COMMITMENT

CHRONOPACKET converts fixed evidence access into a bounded commitment before generation. Its technical core is a typed set objective in which directional supersession and legitimate coexistence are different operations whose effect depends on query intent. The pipeline comprises claim extraction, typed edge prediction, intent classification, and budgeted packet optimization.

For claim extraction, a DeBERTa-v3-large model (He et al., 2023) is fine-tuned using version diffs and release notes together with policy revisions and Wikipedia edit histories. On a human-annotated held-out subset comprising 2,000 labeled claims (Cohen's $\kappa = 0.78$), the extractor attains 81.3 ± 0.7 F1 and $76.8 \pm 1.1\%$ accuracy for validity intervals. Pair blocking retains claim pairs that share a normalized subject–relation key or an explicit source-revision link; this includes cross-passage and within-passage pairs, and each retained claim pair is scored once. A cross-encoder types the retained pairs as

$$\mathcal{E} = \{(c_i, c_j, \tau_{ij}) \mid \tau_{ij} \in \{\textsc{Support}, \textsc{Supersede}, \textsc{Contradict}, \textsc{PartialConflict}, \textsc{Coexist}\}\} \tag{1}$$

with F1 = 84.2 ± 0.6. SUPERSEDE edges are ordered from predecessor to successor; every edge carries weight $w_{ij} \in [0, 1]$, the classifier posterior of its predicted type. Temporal intent $\hat{\iota} \in \mathcal{I}$ is predicted by a two-layer MLP, whose development accuracy is $89.2 \pm 0.8\%$ over the five classes.

Typed packet objective. For each passage, $s_{\mathrm{rel}}(q, p_i)$ is its min–max normalized Contriever-v2 score within the fixed pool. The binary variable $z_{\mathcal{S}}(c)$ is one exactly when the passage containing claim c belongs to packet $\mathcal{S}$; hence all claims in a selected multi-claim passage share its retention decision. For edge $e = (c_i, c_j, \tau_{ij})$, let $r_{\mathcal{S}}(e) = (z_{\mathcal{S}}(c_i), z_{\mathcal{S}}(c_j))$. Query intent and scope define sets $A_\iota(e)$ of compatible retention patterns and $U_\iota(e)$ of unresolved patterns. The edge contributions are

$$\phi_\iota(e, z_{\mathcal{S}}) = w_{ij}\mathbf{1}[r_{\mathcal{S}}(e) \in A_\iota(e)], \qquad \psi_\iota(e, z_{\mathcal{S}}) = \lambda_e w_{ij}\mathbf{1}[r_{\mathcal{S}}(e) \in U_\iota(e)],$$

where λ_e is one except for PARTIALCONFLICT, for which it is the fraction of overlapping claim fields. For a directional SUPERSEDE edge, state-seeking intent rewards the query-valid endpoint alone and penalizes a stale endpoint alone or both incompatible endpoints; CHANGEOVERTIME and VERSIONCOMPARE reward both requested endpoints. SUPPORT rewards both in-scope endpoints. CONTRADICT penalizes retaining both unless DISPUTED or comparison intent explicitly requests both positions. COEXIST rewards exactly the branch set named by query qualifiers. Appendix Table S2 enumerates the patterns, including out-of-scope and no-endpoint cases.

Let $\Pi_\iota(\mathcal{S})$ be deterministic typed pruning: it removes a passage whose claims are all out of query scope or fully superseded by an in-scope endpoint with higher extraction confidence ρ, while protecting endpoints explicitly requested by transition, comparison, or dispute intent; confidence ties use frozen retrieval rank. Pruning is idempotent, $\Pi_\iota(\Pi_\iota(\mathcal{S})) = \Pi_\iota(\mathcal{S})$. Let $\mathcal{E}_{\mathcal{S}}$ contain candidate-pool edges with at least one retained endpoint. We normalize both edge sums by $\max(1, |\mathcal{E}_{\mathcal{S}}|)$, so a packet with no incident scored edge has zero coherence and conflict contribution:

$$\mathrm{Coh}_\iota(\mathcal{S}) = \frac{\sum_{e \in \mathcal{E}_{\mathcal{S}}} \phi_\iota(e, z_{\mathcal{S}})}{\max(1, |\mathcal{E}_{\mathcal{S}}|)}, \qquad \mathrm{Con}_\iota(\mathcal{S}) = \frac{\sum_{e \in \mathcal{E}_{\mathcal{S}}} \psi_\iota(e, z_{\mathcal{S}})}{\max(1, |\mathcal{E}_{\mathcal{S}}|)}.$$

The exact packet objective is

$$\mathcal{S}^\star = \arg \max_{\substack{\mathcal{S} \subseteq \mathcal{P},\ 1 \le |\mathcal{S}| \le B \\ |\Pi_{\hat{\iota}}(\mathcal{S})| \ge 1}} \frac{\sum_{p_i \in \Pi_{\hat{\iota}}(\mathcal{S})} s_{\mathrm{rel}}(q, p_i)}{|\Pi_{\hat{\iota}}(\mathcal{S})|} + \alpha\, \mathrm{Coh}_{\hat{\iota}}(\Pi_{\hat{\iota}}(\mathcal{S})) - \beta\, \mathrm{Con}_{\hat{\iota}}(\Pi_{\hat{\iota}}(\mathcal{S})). \tag{2}$$

The exact committed packet is $\Pi_{\hat{\iota}}(\mathcal{S}^\star)$. Let $J(q, \mathcal{S})$ denote the scored expression in Equation 2. The deployed packet $\widehat{\mathcal{S}}$ is a greedy heuristic for this objective, not the exact argmax. It initializes with the highest-relevance singleton whose pruned image is nonempty. For at most $B - 1$ further insertion rounds, it chooses an unvisited passage with maximum positive post-pruning marginal gain

$$\Delta J(p \mid \mathcal{S}) = J(q, \Pi_{\hat{\iota}}(\mathcal{S} \cup \{p\})) - J(q, \mathcal{S}),$$

then replaces the packet by $\Pi_{\hat{\iota}}(\mathcal{S} \cup \{p\})$. If no singleton survives pruning, the selector returns the empty packet and abstains. Marginal-gain ties use relevance, then mean claim confidence ρ, then frozen retrieval rank. With m claims per passage and $|\mathcal{E}|$ retained claim edges, extraction and edge

construction cost $O(km + |\mathcal{E}|)$ model evaluations after pair blocking, while cached marginal updates cost $O(B(k + |\mathcal{E}|))$. This is appropriate for $k = 20$, $B = 4$, and the observed 2.4 claims per passage. We set $\alpha = 0.6$ and $\beta = 1.2$ on development data. The exact-search control in Table S1 evaluates all $\sum_{b=1}^{4} \binom{20}{b} = 6{,}195$ nonempty budget-feasible pre-pruning packets per query, discards candidates with an empty pruned image, and reports its paired gain and runtime against greedy search on the same fixed subset.

Because every term in J is normalized by packet size or scored-edge count, we temperature-scale $J(q, \widehat{\mathcal{S}})$ on the development split. The binary calibration target is $\mathbf{1}[\widehat{\mathcal{S}} \in \mathfrak{G}(q)]$. Temperature T minimizes development binary negative log likelihood, and τ_{abs} minimizes packet-error risk at the pre-specified development coverage target; both are then frozen before test evaluation. The selector returns $\perp$ for an empty packet or when $\sigma(J/T) < \tau_{\text{abs}}$, so abstention depends on the final packet rather than raw reader confidence.

Intent-conditioned commitment reversal. For a query scoped to Python 3.8, the frozen pool contains a 3.8 tutorial page supporting the requested behavior and a newer Python 3.12 deprecation notice. Recency Rerank chooses the notice, and Reader-Conditioned Selector selects the same out-of-scope endpoint in this trace. CHRONOPACKET extracts the shared subject–relation pair, predicts directional supersession for current Python, detects the 3.8 scope mismatch, and commits to the tutorial. For a comparison query over the same pool, VERSIONCOMPARE intent converts the edge into a compatible transition, retains both endpoints, and enables branch-correct citations. One typed relation thus induces opposite packet decisions under different intents, a set operation unavailable to generic contradiction removal.

Algorithm 1 CHRONOPACKET: query-conditioned temporal packet selection

Require: Query q, retrieved passages $\mathcal{P} = \{p_1, \ldots, p_k\}$
Compute fixed-pool relevance scores $r_i \leftarrow s_{\text{rel}}(q, p_i)$
Extract claims: $\mathcal{C}_i \leftarrow \text{ClaimExtractor}(p_i)$ for each p_i
Block subject–relation/revision pairs and build $\mathcal{G} = (\bigcup_i \mathcal{C}_i, \mathcal{E})$
Classify intent: $\hat{\iota} \leftarrow \text{IntentClassifier}(q)$
Initialize $\widehat{\mathcal{S}}$ with the highest-r_i singleton surviving $\Pi_{\hat{\iota}}$
if no singleton survives typed pruning **then**
  return $\perp$
end if
Mark the initial passage visited; set insertion round $b \leftarrow 1$
while $b < B$ and some unvisited passage yields positive post-pruning ΔJ **do**
  Choose unvisited p with largest $\Delta J(p \mid \widehat{\mathcal{S}})$
  $\widehat{\mathcal{S}} \leftarrow \Pi_{\hat{\iota}}(\widehat{\mathcal{S}} \cup \{p\})$; mark p visited
  $b \leftarrow b + 1$
end while
if $\sigma(J(q, \widehat{\mathcal{S}})/T) < \tau_{\text{abs}}$ **then**
  return $\perp$
end if
return $\widehat{\mathcal{S}}$ to reader

Training targets and isolation. The claim extractor learns span boundaries, normalized subject–relation–object fields, qualifiers, and validity intervals from aligned revision pairs. The edge classifier uses labeled revision links as positive SUPERSEDE/PARTIALCONFLICT targets, same-fact compatible snapshots as SUPPORT/COEXIST targets, and hard negatives matched on subject and relation but differing in object, qualifier, or interval. The intent classifier is trained from query templates and independently labeled natural queries. Query IDs and snapshot groups are disjoint across train, development, and test; source-family-disjoint evaluation is reported separately in Table 3. The three components are trained independently and frozen before packet selection, preventing reader-specific end-to-end co-adaptation.

5 EXPERIMENTS

5.1 BENCHMARK, METRICS, AND MATCHED POLICIES

Source histories and query construction. We evaluate three evolving corpora: API Docs (Python standard-library versions 3.8–3.12; 48K passages and 3.2K changed facts), Policy Revisions (US federal regulation revisions from 2020–2024; 62K passages and 1.8K changes), and Entity Status (Wikipedia infobox snapshots; 85K passages and 4.1K status changes). For each sampled revision chain, the annotation process fixes the target time, version, or branch; writes the query and accepted answer variants; marks a complete minimal support set; and records claim-validity intervals and scope. Retrieval then forms the top-20 pool, and an example enters the controlled study only when at least one complete gold set of at most $B = 4$ passages is contained in that pool. Query IDs and source-snapshot groups are split before feature extraction. The resulting 1800 queries are divided into 960/240/600 train/development/test examples.

Stress strata and annotation. The test set contains 360 Standard cases, where recency agrees with the target, and 240 MR cases, where timestamp-first commitment fails despite complete gold access. MR membership therefore defines a recency-failure stress stratum. Two annotators independently label gold support on 180 queries and MR structure on 120 queries; disagreements are resolved against the source revision chain before model evaluation. Structural labels use the fixed precedence Partial-Rollback, Branch-Specific, Missing-Timestamp, then Wrong-Newer, yielding 48, 60, 60, and 72 test cases, respectively. Appendix Table S3 gives the complete denominator ledger. Aggregate accuracy in Table 1 is the pooled micro-average over the named stratum. To avoid unstable domain cells, topology diagnostics in Table 2 are per-seed domain-macro means over cells with at least 12 examples; the caption identifies the included domains, and Table S9 reports aggregate per-domain outcomes.

Outcome semantics and uncertainty. Answer accuracy is exact match after case, punctuation, whitespace, numeric-format, and pre-registered alias normalization; multi-field and comparison questions require every requested field. Abstentions count as incorrect in all-query accuracy. Citations are mapped to annotated claims in their source snapshot. A query has a stale citation if any cited claim used as answer support has an interval or branch incompatible with the target; a mixed packet is therefore stale when it contains such cited support. Unresolved-conflict incidence is one when the packet delivered to the reader retains any annotated incompatible claim pair not requested by comparison or dispute intent; it is computed from gold claim relations, independent of CHRONOPACKET's predicted graph. For Long-Context Reader the delivered packet is all 20 passages. All-query rates treat abstention as no returned packet, answered-query rates condition on a returned answer, and coverage is the fraction answered. Appendix Table S14 reports both denominators and the no-abstention control. Five seeds vary learned-component initialization and training order while pools, splits, reader prompts, and greedy decoding remain fixed. Tables report the mean and standard deviation across seeds; paired 95% intervals resample shared query IDs and take percentile bounds on the seed-averaged paired difference. The mean conflict-trigger population is 208 of 600 queries (34.7%).

Matched commitment policies. Every method receives the same frozen Contriever-v2 (Izacard et al., 2022) pool and Llama-3.1-70B-Instruct reader (Grattafiori et al., 2024). Packet methods return at most $B = 4$ passages; Long-Context Reader consumes all 20. Recency Rerank instantiates h_{time}, NLI Filter uses generic contradiction scores without temporal edge types, and Reader Self-Rerank uses reader-assigned passage utility. Cross-Encoder Selector and Reader-Conditioned Selector train on the same 960-query partition, use gold-packet membership as the passage-utility target with within-pool hard negatives, tune on the same 240-query development set, and select the top-B passages without interval labels or typed edges. The parameter-matched graph-free control in Table S1 uses the same training queries and packet budget as CHRONOPACKET. These baselines always answer; only CHRONOPACKET applies its calibrated packet-level abstention. Prompts, citation format, maximum 128 generated tokens, greedy decoding, and batch size one are fixed. Training uses 8×A100 80GB GPUs, CUDA 12.4, PyTorch 2.3.1, and bfloat16; extractor, edge, and intent training require 2.0, 1.5, and 0.5 hours. Reader inference uses one 8×A100 80GB node with tensor parallelism. Latency is post-retrieval time to first token over answered queries after 50 warm-up queries, excluding autoregressive

decoding shared by all methods; selector-only latency, including abstentions, is reported separately. Independence checks remove overlap unions, hold out source families or entities, and repeat transfer with BGE-M3 (Chen et al., 2024); the incident-update extension adds a faster, less-structured source family.

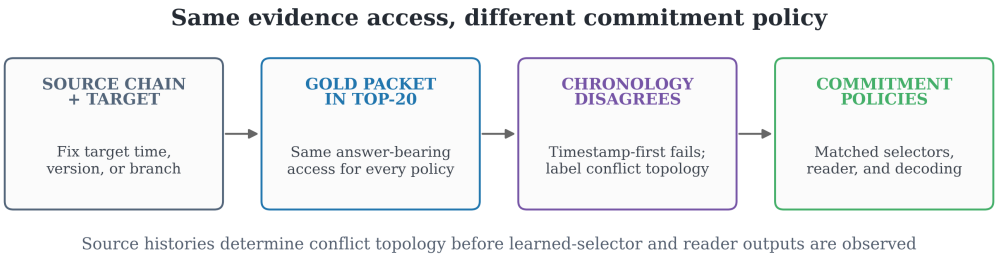

Figure 1: Access–commitment protocol. Timestamp-first failure defines the MR stress stratum; source histories then determine the four structural labels before learned-policy outputs are observed. Gold-containing pools, reader, prompt, and decoding are frozen while the commitment rule varies.

Figure 1 makes the conditioning explicit: construction first identifies recency failures with guaranteed access, then the matched evaluation localizes policy differences to the commitment stage. The experiments answer three questions: whether typed packets improve answers and support under identical access; which conflict topologies and edge operations produce the change; and whether the ordering survives independent sources, entities, retrievers, and update cadence.

5.2 Results

Table 1 compares commitment policies under identical evidence access. MR and Wrong-Newer accuracy measure whether the committed evidence yields the scoped answer; stale citations and unresolved contradictions measure the reliability of its support packet.

Table 1: Evidence commitment under matched access: Contriever-v2, $k = 20$, and Llama-3.1-70B-Instruct. Accuracy and all-query reliability columns are pooled micro-averages over the named stratum; abstentions count as incorrect for accuracy. Entries are mean ± std over five seeds; Cov. is answer coverage and p50 is answered-query post-retrieval time to first token in ms.

Method	Std. ↑	MR ↑	WN ↑	Stale ↓	Contr. ↓	Cov.	p50
Recency Rerank	78.4±0.6	52.1±1.1	43.7±1.4	18.8±0.7	28.6±0.8	100	42
NLI Filter	75.2±0.7	58.7±0.9	51.3±1.1	15.2±0.6	18.4±0.7	100	67
Reader Self-Rerank	77.3±0.6	56.8±1.0	49.7±1.2	16.3±0.6	19.3±0.7	100	156
Long-Context Reader	80.1±0.5	61.3±0.8	54.2±1.0	13.2±0.5	22.1±0.7	100	284
Cross-Encoder Selector	80.8±0.5	64.7±0.9	58.4±1.0	9.6±0.4	11.8±0.6	100	109
Reader-Conditioned Selector	81.0±0.5	65.1±0.8	58.9±0.9	9.1±0.4	10.7±0.6	100	171
ChronoPacket	*81.6±0.5*	*66.4±0.8*	*60.1±0.9*	*7.8±0.4*	*6.3±0.5*	93.8±0.4	61

Under the matched graph-free comparison, ChronoPacket improves over Reader-Conditioned Selector by 1.3 MR points (paired 95% CI [0.4, 2.3]), 1.2 Wrong-Newer points, 1.3 stale-citation points, 4.4 contradiction points, and 110 ms answered-query p50 latency. Relative to Long-Context Reader, the corresponding MR, stale-citation, and contradiction differences are 5.1, 5.4, and 15.8 points. The no-abstention control reaches 76.6% overall accuracy and 8.1% stale citations, compared with 74.7% and 9.1% for Reader-Conditioned Selector (Table S14), preserving the overall accuracy and citation ordering without selective prediction. At the selected 93.8%-coverage operating point, answered accuracy is 80.5%. Relative to Long-Context Reader, the 1.5-point Standard difference concentrates the larger effect on cases where identical access permits incompatible commitments.

The topology diagnostics cover every MR structure: relative to Long-Context Reader, ChronoPacket gains 5.9 Wrong-Newer, 4.6 Missing-Timestamp, 6.3 Branch-Specific, and 4.1

Table 2: Topology diagnostics and mechanism signatures. Panel A is a per-seed domain-macro mean over cells with at least 12 examples: Wrong-Newer uses all three domains; Missing-Timestamp uses Policy Revisions and Entity Status; Branch-Specific and Partial-Rollback use API Docs and Entity Status. Panel B changes one typed operation while retrieval, reader, packet budget, supervision, and decoding remain fixed. Panel C reports citation behavior on the two topologies with the largest answer gains over full-pool reading.

Method	Wrong-Newer	Missing-TS	Branch	Rollback
Recency Rerank	43.7±1.4	49.4±1.2	58.2±1.0	55.8±1.1
NLI Filter	51.3±1.1	56.8±1.0	63.4±0.9	61.2±1.0
Long-Context Reader	54.2±1.0	59.7±0.9	66.1±0.8	63.8±0.9
CHRONOPACKET	*60.1±0.9*	*64.3±0.8*	*72.4±0.7*	*67.9±0.8*

Variant	MR	Wrong-Newer	Branch	Rollback
CHRONOPACKET	*66.4±0.8*	*60.1±0.9*	*72.4±0.7*	*67.9±0.8*
Supersession → contradiction	61.4±0.9	53.8±1.2	68.9±0.9	61.6±1.0
Coexistence → non-conflict	62.9±0.9	58.5±1.0	63.7±1.0	66.2±0.9
Shuffled intervals, matched density	59.8±1.0	51.7±1.3	65.1±1.0	61.0±1.1

Stale-citation slice (%)	n	Long-Context	CHRONOPACKET
Wrong-Newer	72	17.8±0.7	10.2±0.5
Branch-Specific	60	8.6±0.5	4.9±0.3

Rollback points (Table 2). The interventions expose distinct signatures. Recasting supersession as generic contradiction removes 6.3 Wrong-Newer and 6.3 Rollback points; collapsing coexistence removes 8.7 Branch-Specific points; shuffling intervals removes 6.6 aggregate MR points. Wrong-Newer and Branch-Specific stale-citation rates fall by 7.6 and 3.7 points relative to Long-Context Reader. Reader-Conditioned Selector narrows the answer gap to 1.2 and 2.0 points on those two slices (Table S11), while the larger contradiction reductions in Table 1 show what the typed objective changes in the committed support.

Table 3: Independence checks under frozen retrieval. Panel A reports MR accuracy; Δ is CHRONOPACKET minus Reader-Conditioned Selector with a paired-bootstrap 95% CI. Panel B reports all-query stale-citation incidence, answered-query stale rate for CHRONOPACKET, and answer coverage. All point estimates are mean ± std over five seeds.

Condition	n	n_{MR}	Long-Ctx.	X-Enc.	Reader-Cond.	CHRONOPACKET	Δ [95% CI]
Full test	600	240	61.3±0.8	64.7±0.9	65.1±0.8	66.4±0.8	1.3 [0.4, 2.3]
Union-clean	503	202	61.1±0.8	64.2±0.9	64.6±0.9	65.9±0.9	1.3 [0.3, 2.5]
Source holdout	418	168	58.7±1.0	61.5±1.1	62.3±1.0	64.2±0.9	1.9 [0.7, 3.2]
Entity disjoint	451	180	59.4±1.0	62.4±1.0	63.0±0.9	64.8±0.9	1.8 [0.6, 3.1]
External transfer	180	72	59.1±1.2	61.7±1.2	62.4±1.1	63.8±1.0	1.4 [0.2, 2.8]
Incident updates	240	108	55.8±1.3	59.4±1.2	60.2±1.2	63.1±1.1	2.9 [1.4, 4.5]
Incidents, BGE-M3	240	108	57.0±1.2	—	61.0±1.1	63.8±1.0	2.8 [1.3, 4.4]

Panel B: citation reliability and coverage (%)

Condition	Long-Ctx. stale	Reader-Cond. stale	CP stale	CP answered stale	CP coverage
Full test	13.2±0.5	9.1±0.4	7.8±0.4	8.3±0.4	93.8±0.4
Union-clean	13.3±0.5	9.4±0.5	8.0±0.4	8.5±0.5	93.6±0.5
Source holdout	14.4±0.6	10.3±0.5	8.7±0.5	9.2±0.5	93.1±0.6
Entity disjoint	14.0±0.6	10.0±0.5	8.4±0.5	8.9±0.5	93.5±0.5
External transfer	14.6±0.6	10.8±0.6	9.4±0.5	9.9±0.5	94.4±0.6
Incident updates	15.8±0.7	10.8±0.6	8.9±0.6	9.6±0.6	92.4±0.7
Incidents, BGE-M3	15.2±0.7	10.4±0.6	8.5±0.5	9.2±0.6	92.7±0.7

Excluding the union of audited source, entity-family, and near-duplicate overlap leaves 503 queries and preserves the MR ordering. Source-family and entity holdouts retain CHRONOPACKET's 1.9-

and 1.8-point advantages over Reader-Conditioned Selector. The separately adjudicated 180-query transfer set changes the construction path, while 240 incident/status-update queries accelerate and disorder the revision cadence; Contriever-v2 and BGE-M3 preserve the ranking. At 92.4–94.4% coverage, answered-query stale rates remain below matched alternatives, establishing a citation benefit beyond selective abstention.

5.3 Mechanism Analysis

Controlled interventions map performance to distinct commitment operations (Appendix Table S1). Collapsing supersession produces the largest losses on Wrong-Newer and Rollback; collapsing coexistence removes 8.7 Branch-Specific points. Shuffling intervals at matched graph density removes 6.6 MR points, and the parameter-matched graph-free scorer trails the typed objective by 2.6 points. Exact search adds only 0.7 MR points on the same 120-query subset (95% CI [0.2, 1.3]), locating the advantage in temporal structure instead of a larger combinatorial search budget. Appendix Figure S1 traces three corresponding packet decisions; Tables S4 and S5 report component quality and broader removals.

6 Discussion

The access–commitment decomposition changes what temporal RAG evaluation can attribute. When retrieval and selection move together, a correct answer may reflect better access, better evidence use, or both. Gold-containing frozen pools and answer-plus-citation metrics identify the second contribution inside a recency-conditioned regime. The advantage over learned graph-free selectors survives overlap exclusion, source-family and entity holdouts, external construction, faster incident updates, and BGE-M3 retrieval; targeted removals then attribute the gain to interval semantics, directional supersession, coexistence, and query intent rather than packet budget or search capacity.

The evidence establishes this ordering for English evolving corpora, gold-containing top-k pools, the evaluated reader family, and predicted claim, interval, edge, and intent structure. MR measures behavior conditional on a recency failure, while Standard results cover the complementary agreement regime; together they characterize commitment across the study mixture. The remaining 33.6% MR error and 8.3% answered-query stale-citation rate concentrate the next technical opportunities in interval and relation estimation and citation verification. Source-family, entity-disjoint, transfer, second-retriever, and fast-update results show that the measured ordering extends across templates, entities, retrieval models, and revision cadence.

7 Conclusion

Publication chronology and query-compatible evidence need not select the same revision. The MR protocol makes that disagreement measurable under guaranteed access, and ChronoPacket acts on it through intent-conditioned supersession, coexistence, and rollback semantics over compact evidence packets. Under identical pools and reader, these operations improve answers, sharply reduce unresolved support conflicts, and preserve their ordering across independent source, entity, retriever, and update-cadence checks. Temporal RAG systems should therefore be evaluated on both access to an answer-bearing snapshot and the coherent claim set that actually governs generation.

AI Use Statement

Generative AI tools assisted manuscript preparation, including editorial revision, formatting and consistency checks, and bibliographic verification. Responsibility for the final manuscript, including its scientific content, results, and claims, rests with the authors.

References

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*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Jiawei Cao, Jie Ouyang, Mingyue Cheng, Zhaomeng Zhou, Chunli Liu, Yupeng Li, Zirui Liu, and Shijin Wang. Re3: Relevance & recency retrieval for mitigating temporal hallucination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 25735–25760, 2026. doi: 10.18653/v1/2026.acl-long.1180. URL https://aclanthology.org/2026.acl-long.1180/.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. doi: 10.1162/tacl_a_00459. URL https://aclanthology.org/2022.tacl-1.15/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, 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, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of ICML*, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Jiale Han, Austin Cheung, Yubai Wei, Zheng Yu, Xusheng Wang, Bing Zhu, and Yi Yang. RAG meets temporal graphs: Time-sensitive modeling and retrieval for evolving knowledge. *arXiv preprint arXiv:2510.13590*, 2025. doi: 10.48550/arXiv.2510.13590. URL https://arxiv.org/abs/2510.13590.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024. doi: 10.18653/v1/2024.findings-naacl.159. URL https://aclanthology.org/2024.findings-naacl.159/.

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*, 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. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Chaeyun Jang, Deukhwan Cho, Seanie Lee, Hyungi Lee, and Juho Lee. Reliable decision-making via calibration-oriented retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Rgk129n73h.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023. doi: 10.48550/arXiv.2207.13332. URL https://arxiv.org/abs/2207.13332.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 17606–17619, 2025. doi: 10.18653/v1/2025.acl-long.861. URL https://aclanthology.org/2025.acl-long.861/.

Dong Li, Yichen Niu, Ying Ai, Xiang Zou, Biqing Qi, and Jianxing Liu. T-GRAG: A dynamic GraphRAG framework for resolving temporal conflicts and redundancy in knowledge retrieval. *arXiv preprint arXiv:2508.01680*, 2025. doi: 10.48550/arXiv.2508.01680. URL https://arxiv.org/abs/2508.01680.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, et al. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of ICML*, 2022. doi: 10.48550/arXiv.2205.11388. URL https://arxiv.org/abs/2205.11388.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Xi Nie, Hongwei Li, Shenghao Wu, Wenshu Fan, Qiyang Song, and Wenbo Jiang. EvoTrustRAG: Evolution-aware conflict attribution and evidence handling for reliable retrieval-augmented generation. *arXiv preprint arXiv:2608.07933*, 2026. doi: 10.48550/arXiv.2608.07933. URL https://arxiv.org/abs/2608.07933.

Qingying Niu, Yuhao Wang, Ruiyang Ren, Bohui Fang, and Wayne Xin Zhao. ArbGraph: Conflict-aware evidence arbitration for reliable long-form retrieval-augmented generation. *arXiv preprint arXiv:2604.18362*, 2026. doi: 10.48550/arXiv.2604.18362. URL https://arxiv.org/abs/2604.18362.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. doi: 10.48550/arXiv.1901.04085. URL https://arxiv.org/abs/1901.04085.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605. URL https://aclanthology.org/2023.tacl-1.75/.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. ReliabilityRAG: Effective and provably robust defense for RAG-based web-search. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1522. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/41457d56a2fdd20b4c072190290a549b-Abstract-Conference.html.

Xiaonan Si, Meilin Zhu, Simeng Qin, Lijia Yu, Lijun Zhang, Shuaitong Liu, Xinfeng Li, Ranjie Duan, Yang Liu, and Xiaojun Jia. SeCon-RAG: A two-stage semantic filtering and conflict-free framework for trustworthy RAG. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-2376. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/668563ef18fbfef0b66af491ea334d5f-Abstract-Conference.html.

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. URL https://aclanthology.org/2024.findings-acl.813/.

Chenyu Wang, Yingmin Liu, and Yang Shu. ConflictRAG: Detecting and resolving knowledge conflicts in retrieval augmented generation. *arXiv preprint arXiv:2605.17301*, 2026a. doi: 10.48550/arXiv.2605.17301. URL https://arxiv.org/abs/2605.17301.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 30553–30571, 2025a. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *Conference on Language Modeling*, 2025b. URL https://arxiv.org/abs/2504.13079.

Xiaoman Wang, Yaoze Zhang, Wenzhuo Fan, Hongwei Zhang, Ding Wang, Guohang Yan, Song Mao, Botian Shi, Yunshi Lan, and Pinlong Cai. IA-RAG: Interval-algebra-driven temporal reasoning for dynamic knowledge retrieval. *arXiv preprint arXiv:2606.06044*, 2026b. doi: 10.48550/arXiv.2606.06044. URL https://arxiv.org/abs/2606.06044.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024. doi: 10.48550/arXiv.2405.15556. URL https://arxiv.org/abs/2405.15556.

Neeraj Yadav. Temporal validity in retrieval memory: Eliminating stale-fact errors for AI agents over evolving knowledge. *arXiv preprint arXiv:2606.26511*, 2026. doi: 10.48550/arXiv.2606.26511. URL https://arxiv.org/abs/2606.26511.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=S1fc92uemC.

Linda Zeng, Rithwik Gupta, Divij Motwani, Yi Zhang, and Diji Yang. Worse than zero-shot? a fact-checking dataset for evaluating the robustness of RAG against misleading retrievals. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ed25c00ff6900989116d3ba5d607d33d-Abstract-Datasets_and_Benchmarks_Track.html.

Michael J. Q. Zhang and Eunsol Choi. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.586. URL https://aclanthology.org/2021.emnlp-main.586/.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 21863–21882, 2025. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062/.

ADDITIONAL EXPERIMENTAL ANALYSIS

Figure S1 connects the aggregate mechanism results to three individual decisions. The traces show how the same shared-pool comparison becomes an out-of-scope rejection, a branch selection after preserving coexistence, or a directed rollback traversal.



Figure S1: Three matched-access traces expose distinct commitment operations. CHRONOPACKET rejects an out-of-scope newer revision, preserves branch coexistence before selecting the named branch, and follows a rollback chain to the restored state. In each case, packet choice determines both answer scope and citation support.

Table S1: Controlled mechanism analysis with retrieval, reader, packet budget, supervision, decoding, and five seeds fixed. Panel A reports hard-slice accuracy; Panel B reports all-query packet-reliability incidence and systems behavior. Greedy and exact search use the same fixed 120-query subset; exact search improves paired MR accuracy by 0.7 points (95% CI [0.2, 1.3]).

Variant	MR ↑	Wrong-Newer ↑	Branch ↑	Rollback ↑
CHRONOPACKET	66.4±0.8	60.1±0.9	72.4±0.7	67.9±0.8
Untyped conflict edges	63.0±0.9	55.9±1.1	68.5±0.9	65.0±1.0
Supersession → contradiction	61.4±0.9	53.8±1.2	68.9±0.9	61.6±1.0
Coexistence → non-conflict	62.9±0.9	58.5±1.0	63.7±1.0	66.2±0.9
Shuffled intervals, matched density	59.8±1.0	51.7±1.3	65.1±1.0	61.0±1.1
Intent-agnostic coherence	61.8±0.9	57.2±1.1	65.9±0.9	62.4±1.0
Parameter-matched graph-free	63.8±0.9	57.3±1.1	68.4±0.9	65.1±1.0
Greedy packet search (subset)	66.4±0.9	60.0±1.1	72.1±1.0	68.0±1.1
Exact packet search (subset)	67.1±0.8	60.8±0.9	73.2±0.8	68.4±0.9

Variant	Stale ↓	Contradiction ↓	Packet size	Coverage	p50 ms
CHRONOPACKET	7.8±0.4	6.3±0.5	2.8±0.1	93.8±0.4	61
Untyped conflict edges	9.5±0.5	11.4±0.6	2.9±0.1	93.6±0.5	58
Supersession → contradiction	10.8±0.5	12.7±0.7	2.9±0.1	93.5±0.5	56
Coexistence → non-conflict	9.2±0.5	13.0±0.7	2.6±0.1	93.7±0.5	57
Shuffled intervals, matched density	12.2±0.6	15.3±0.7	2.8±0.1	93.4±0.6	59
Intent-agnostic coherence	8.7±0.5	9.4±0.6	2.8±0.1	93.7±0.5	59
Parameter-matched graph-free	10.0±0.5	11.6±0.6	3.0±0.1	94.0±0.5	64
Greedy packet search (subset)	7.9±0.5	6.4±0.6	2.8±0.1	93.6±0.5	61
Exact packet search (subset)	7.6±0.4	6.1±0.5	2.8±0.1	93.8±0.5	183

The intervention effects are operation-specific. Supersession primarily protects Wrong-Newer and Rollback decisions, coexistence protects branch selection, and interval shuffling degrades all hard regimes without changing graph density; the modest exact-search gain shows that these effects do not reduce to search capacity.

Table S2: Complete typed-edge retention semantics. Patterns are (z_i, z_j) after query-scope resolution; for SUPERSEDE, i is the predecessor and j the successor. "Requested" means explicitly named by the query's time, version, branch, transition, or dispute scope. Compatible patterns contribute w_{ij} to ϕ; unresolved patterns contribute $\lambda_e w_{ij}$ to ψ.

Edge type	State-query compatible	Transition/comparison compatible	Dispute-query compatible	Unresolved pattern
SUPPORT	$(1,1)$ for two in-scope supporting endpoints	$(1,1)$ when both endpoints support the requested transition	$(1,1)$ for requested common support	None
SUPERSEDE	$(0,1)$ when j is query-valid; $(1,0)$ when the query targets i's historical interval	$(1,1)$ when both endpoints define the requested transition	Exact requested endpoint set	Stale endpoint alone, or both endpoints when the pair is not requested
CONTRADICT	The single endpoint selected by time and scope	$(1,1)$ only when the query requests the contrast	$(1,1)$ when both positions are requested	$(1,1)$ whenever both claims remain but the query does not request both
PARTIALCONFLICT	Same endpoint rule as CONTRADICT	Same requested-pair rule as CONTRADICT	Same requested-pair rule as CONTRADICT	Same as CONTRADICT, scaled by overlapping-field fraction λ_e
COEXIST	Singleton named by branch/scope	$(1,1)$ when both named branches or versions are compared	Exact requested position set	None; unrequested endpoints receive no reward

Pattern $(0,0)$ always contributes zero. An endpoint excluded by query scope is unrequested. Because passage selection jointly retains all claims in a passage, any incident edge forced into an unresolved pattern is penalized once; duplicate passage pairs remain distinct only when they connect distinct claim pairs.

Table S2 makes the selector's decision rule auditable at the edge level: query scope determines which endpoint pattern is compatible, while retained but unrequested conflict receives an explicit penalty. Table S3 complements that semantic specification with the exact populations and aggregation rule used by every headline result.

Table S3: Dataset composition and protocol accounting. Table 1 reports pooled micro-averages over the named query set; Table 2 treats each topology as a separate diagnostic estimand, and Table S9 reports domains separately.

Domain	Test	Std.	MR	WN	MTS	BS	PR
API Docs	200	120	80	18	10	34	18
Policy Revisions	200	136	64	28	24	4	8
Entity Status	200	104	96	26	26	22	22
Total	600	360	240	72	60	60	48

Design quantity	Value	Experimental role
Total labeled queries / train / dev / test	1800 / 960 / 240 / 600	Fixed partitions and denominators for all reported experiments
Packet budget B / gold-label dual annotation / MR-label dual annotation	4 / 180 / 120	Fixed selector budget and independently double-annotated evaluation subsets

Aggregation note. Each seed computes query-level indicators before averaging. Headline MR and Standard values weight every query in the named pooled stratum equally. Topology diagnostics are domain-macro means over cells with at least 12 examples: Wrong-Newer includes all domains, Missing-Timestamp includes Policy Revisions and Entity Status, and Branch-Specific and Partial-Rollback include API Docs and Entity Status. They are reported as conditional diagnostics, not as an inversion of the pooled MR estimate.

Table S3 makes every headline denominator explicit. The pooled estimand weights queries equally, while the per-domain table reveals how those query-level outcomes are distributed across source families.

Table S4: Component quality and selector engagement. The conflict count is the rounded mean count across five seeds; its parenthesized value is the corresponding mean rate ± std.

Component	Metric	Value
Claim Extractor (DeBERTa-v3-large)	F1 on held-out 2K claim instances	81.3 ± 0.7
	Validity interval accuracy	76.8 ± 1.1%
	Average claims per passage	2.4
Edge Classifier (DeBERTa-v3-large)	F1 (5-class, cross-validation)	84.2 ± 0.6
	Average conflict edges per query	3.8
Intent Classifier (2-layer MLP)	Accuracy (5-class)	89.2 ± 0.8%
Selector operation	Queries with ≥1 temporal conflict	208 (34.7 ± 1.3%)
	Abstention trigger rate	6.2%
	Forced-answer error on abstained queries	82.1%
	Selector-only overhead p50 / p95	19 / 23 ms
	Selection-to-first-token p50 / p95	61 / 81 ms

Table S4 shows that the selector engages on roughly one third of queries and that abstention is concentrated on difficult cases: forcing answers on abstained queries yields 82.1% error. Table S5 separates component quality from selective behavior; removing coherence selection produces the largest MR and stale-citation degradation, while Table S14 reports the no-abstention operating point directly.

Split keys are computed before feature extraction. Union-clean removes the union of source, entity-family, and near-duplicate overlap; source-family holdout partitions publisher/template families; entity-disjoint partitions normalized entity identifiers. The extractor, edge model, and intent classifier are retrained on the corresponding training partition, then frozen before the same Contriever-v2 pools, Llama-3.1-70B reader, $B = 4$, decoding, and five-seed evaluation are applied. Paired intervals and win/tie/loss counts use the same MR queries for CHRONOPACKET and Reader-Conditioned Selector.

Table S5: Broad component-removal study. Entries are mean ± std over five seeds except ECE; selective prediction is reported separately in Table S14.

Variant	MR Acc. ↑	Stale Cite (%) ↓	Abstain ECE ↓
CHRONOPACKET (full)	*66.4±0.8*	*7.8±0.4*	*0.04*
− Claim extraction (raw passages)	59.2±0.9	12.1±0.5	0.09
− Temporal graph (pairwise only)	62.1±0.8	9.4±0.4	0.06
− Intent classification	61.8±0.9	8.7±0.5	0.05
− Coherence selection (recency fallback)	55.3±1.0	15.4±0.6	0.12

Table S6: Access–commitment evidence matrix. Each row fixes the accessible evidence, identifies the temporal retention decision, and links its most direct controlled challenge to manuscript evidence.

Temporal case and fixed trust condition	Commitment property and direct controlled check	Evidence
Wrong-Newer answer and citation; a compatible snapshot remains in top-k	Retain the query-valid endpoint despite a later false revision. Accuracy rises 5.9 points over full-pool reading; replacing supersession with generic contradiction removes 6.3 points.	Tables 2, S1
Missing-Timestamp answer and citation; claim content supports interval inference	Recover semantic validity when publication and validity times diverge. Accuracy rises 4.6 points over full-pool reading; interval shuffling removes 6.6 aggregate MR points.	Tables 2, S1
Branch-Specific answer; the query names one branch or a comparison	Preserve coexistence while committing only the requested branch. Accuracy rises 6.3 points over full-pool reading; collapsing coexistence removes 8.7 points.	Tables 2, S1
Partial-Rollback state; the rollback event remains in top-k	Follow the directional chain instead of anchoring an intermediate revision. Accuracy rises 4.1 points over full-pool reading; replacing supersession with contradiction removes 6.3 points.	Tables 2, S1
Method ordering; candidate pools and labels remain frozen	Preserve the ranking after overlap-union exclusion and source-family/entity holdouts against Reader-Conditioned Selector.	Tables 3, S8

The matrix makes the decision boundary explicit: retrieval fixes accessible evidence, and CHRONOPACKET governs commitment inside that pool. Each temporal stressor is paired with the most direct comparator and the held-out condition that tests its behavior.

The next two tables pair the exact population retained by each independence check with its matched-policy outcome. Reading denominators and results together shows that the advantage is preserved after removing overlap and after changing the held-out source or entity unit, rather than arising from a shift in test composition.

Table S7: Pre-evaluation overlap-exclusion and holdout denominators by condition and domain. "Clean" excludes the union of exact-source, entity-family, and near-duplicate snapshot overlap.

Condition	Domain	Total	Standard	MR	Wrong-Newer	Branch
Full	API / Policy / Entity	600	360	240	72	60
Clean	API Docs	165	99	66	15	28
	Policy Revisions	167	114	53	23	3
	Entity Status	171	88	83	23	18
	All domains	503	301	202	61	49
Source-family held out	API Docs	132	80	52	13	22
	Policy Revisions	137	94	43	19	3
	Entity Status	149	76	73	18	17
	All domains	418	250	168	50	42
Entity disjoint	API Docs	145	88	57	14	23
	Policy Revisions	149	102	47	21	3
	Entity Status	157	81	76	19	18
	All domains	451	271	180	54	44

Table S8: Overlap-exclusion and holdout performance over five seeds. The upper panel reports accuracy and paired MR differences against Reader-Conditioned Selector; win/tie/loss is the query-level MR comparison. The lower panel distinguishes all-query from answered-query packet failures.

Condition	Method	Standard ↑	MR ↑	WN ↑	Branch ↑	Δ_{MR} [95% CI]	W/T/L
Union-clean	Long-Context	80.3±0.6	61.1±0.8	54.0±1.2	65.8±0.9	—	—
	Cross-Encoder	80.9±0.5	64.2±0.9	58.2±1.1	69.7±0.9	—	—
	Reader-Conditioned	81.1±0.5	64.6±0.9	58.7±1.0	70.2±0.8	—	—
	CHRONOPACKET	81.7±0.5	65.9±0.9	60.0±1.0	72.1±0.8	1.3 [0.3, 2.5]	23/161/18
Source holdout	Long-Context	79.5±0.6	58.7±1.0	49.8±1.3	64.0±1.0	—	—
	Cross-Encoder	80.1±0.6	61.5±1.1	54.7±1.2	67.5±0.9	—	—
	Reader-Conditioned	80.3±0.6	62.3±1.0	55.4±1.2	68.3±0.9	—	—
	CHRONOPACKET	81.0±0.5	64.2±0.9	58.1±1.1	70.6±0.8	1.9 [0.7, 3.2]	19/133/16
Entity disjoint	Long-Context	79.1±0.6	59.4±1.0	51.0±1.3	64.8±1.0	—	—
	Cross-Encoder	79.9±0.6	62.4±1.0	55.9±1.2	68.0±0.9	—	—
	Reader-Conditioned	80.2±0.6	63.0±0.9	56.5±1.1	68.8±0.9	—	—
	CHRONOPACKET	80.9±0.5	64.8±0.9	59.0±1.1	71.0±0.8	1.8 [0.6, 3.1]	20/143/17

Condition	Method	All/answered stale ↓	All/answered contradiction ↓	Coverage
Union-clean	Long-Context	13.3±0.5 / 13.3±0.5	22.0±0.8 / 22.0±0.8	100.0
	Cross-Encoder	9.8±0.5 / 9.8±0.5	11.6±0.7 / 11.6±0.7	100.0
	Reader-Conditioned	9.4±0.5 / 9.4±0.5	10.5±0.7 / 10.5±0.7	100.0
	CHRONOPACKET	8.0±0.4 / 8.5±0.5	6.2±0.6 / 6.6±0.6	93.6±0.5
Source holdout	Long-Context	14.4±0.6 / 14.4±0.6	23.4±0.8 / 23.4±0.8	100.0
	Cross-Encoder	11.2±0.6 / 11.2±0.6	13.5±0.7 / 13.5±0.7	100.0
	Reader-Conditioned	10.3±0.5 / 10.3±0.5	12.4±0.7 / 12.4±0.7	100.0
	CHRONOPACKET	8.7±0.5 / 9.2±0.5	7.1±0.6 / 7.6±0.6	93.1±0.6
Entity disjoint	Long-Context	14.0±0.6 / 14.0±0.6	22.8±0.8 / 22.8±0.8	100.0
	Cross-Encoder	10.9±0.5 / 10.9±0.5	13.0±0.7 / 13.0±0.7	100.0
	Reader-Conditioned	10.0±0.5 / 10.0±0.5	11.9±0.7 / 11.9±0.7	100.0
	CHRONOPACKET	8.4±0.5 / 8.9±0.5	6.8±0.6 / 7.3±0.6	93.5±0.5

Table S9: Per-domain aggregate results for the main comparison, reported as mean ± std over five seeds. Headline topology diagnostics use the support-threshold aggregation specified in Table S3.

Aggregate accuracy and all-query stale-citation incidence (%)							
Domain	n_{Std}	n_{MR}	LC Std.	CP Std.	LC MR	CP MR	LC / CP stale
API Docs	120	80	81.0 ± 0.6	82.1 ± 0.5	63.5 ± 0.8	68.9 ± 0.7	12.6 ± 0.5 / 7.1 ± 0.4
Policy Revisions	136	64	82.3 ± 0.5	83.5 ± 0.5	58.2 ± 0.9	62.8 ± 0.9	14.8 ± 0.6 / 8.5 ± 0.5
Entity Status	104	96	76.2 ± 0.7	78.5 ± 0.6	61.5 ± 0.9	66.7 ± 0.8	12.2 ± 0.6 / 7.8 ± 0.4

The MR advantage appears in all three domains while Standard differences remain small. Table 2 reports the separate topology diagnostics under the support-threshold rule fixed in Table S3.

Table S10: Retrieval-depth sensitivity on the same query set. This supplementary robustness check changes both accessible candidates and commitment context with k; the fixed-access causal comparison remains the $k = 20$ experiment in Table 1.

Method	k	Std. Acc. ↑	MR Acc. ↑	WN Acc. ↑	Stale Cite (%) ↓	p50 Lat. (ms)
Long-Context Reader	10	79.6 ± 0.6	58.9 ± 0.9	52.4 ± 1.1	13.8 ± 0.5	192
Long-Context Reader	20	80.1 ± 0.5	61.3 ± 0.8	54.2 ± 1.0	13.2 ± 0.5	284
Long-Context Reader	40	79.8 ± 0.6	61.0 ± 0.8	53.8 ± 1.0	13.5 ± 0.5	471
ChronoPacket	10	81.0 ± 0.5	64.8 ± 0.8	58.6 ± 0.9	8.5 ± 0.4	47
ChronoPacket	20	81.6 ± 0.5	66.4 ± 0.8	60.1 ± 0.9	7.8 ± 0.4	61
ChronoPacket	40	81.2 ± 0.5	66.1 ± 0.7	59.6 ± 0.8	7.9 ± 0.4	93

ChronoPacket remains ahead on MR accuracy, Wrong-Newer accuracy, and stale citations at $k \in \{10, 20, 40\}$. This is an access-plus-commitment robustness check: accuracy and latency shift with the available pool for both methods, while the selector ordering stays stable. Retriever-family transfer is measured separately in Table S12.

Table S11 adds two stronger graph-free selectors under exactly the same fixed Contriever-v2 pool, $k = 20$, packet cap $B = 4$, Llama-3.1-70B reader, five-seed protocol, and A100 80GB deployment used in the main paper. Cross-Encoder Selector fine-tunes DeBERTa-v3-large to score passage utility from the frozen pool and then selects the top-B packet without a temporal graph. Reader-Conditioned Selector uses the same reader family to produce passage-level utility scores before top-B selection. Both baselines match the main paper's prompt budget and latency measurement protocol. This makes the table a direct test of whether the fixed-pool separation survives stronger packet scoring rather than a looser comparison across changing retrieval or generation settings.

Table S11: Stronger matched-retrieval selector baselines under the same fixed Contriever-v2 pool, $k = 20$, packet budget $B = 4$, Llama-3.1-70B reader, and five-seed reporting convention. All rows evaluate the same 600 test queries, including 240 MR queries and 60 Branch-Specific MR queries.

Method	Std. Acc. ↑	MR Acc. ↑	WN Acc. ↑	BS Acc. ↑
Long-Context Reader	80.1 ± 0.5	61.3 ± 0.8	54.2 ± 1.0	66.1 ± 0.8
Cross-Encoder Selector	80.8 ± 0.5	64.7 ± 0.9	58.4 ± 1.0	69.8 ± 0.8
Reader-Conditioned Selector	81.0 ± 0.5	65.1 ± 0.8	58.9 ± 0.9	70.4 ± 0.8
ChronoPacket	81.6 ± 0.5	66.4 ± 0.8	60.1 ± 0.9	72.4 ± 0.7

Method	Stale cite ↓	Contradiction ↓	p50 (ms)	p95 (ms)
Long-Context Reader	13.2 ± 0.5	22.1 ± 0.7	284	412
Cross-Encoder Selector	9.6 ± 0.4	11.8 ± 0.6	109	154
Reader-Conditioned Selector	9.1 ± 0.4	10.7 ± 0.6	171	249
ChronoPacket	7.8 ± 0.4	6.3 ± 0.5	61	81

Graph-free learned selectors close much of the gap to Long-Context Reader, establishing competitive packet-utility baselines. Reader-Conditioned Selector performs best among these controls, while

CHRONOPACKET remains higher on MR, Wrong-Newer, and Branch-Specific accuracy and lower on stale citations, contradictions, and latency under the same pool and reader budget.

Table S12 is an independence check built outside the three-domain construction path. It contains 180 release-note, policy-bulletin, and public knowledge-base queries plus 240 incident/status-update queries with faster and less-structured revisions. Annotators first sample a revision chain from a source family absent from benchmark construction, fix the target interval or branch, and write the query before retrieval; Contriever-v2 and BGE-M3 pools are then constructed independently and retained only when each contains a complete compatible support set of at most $B = 4$ passages. Candidate pools are frozen separately under the two retrievers; all methods retain the Llama-3.1-70B-Instruct reader, $k = 20$, $B = 4$, decoding, and five-seed protocol. The original set has 90 dual-annotated queries with 92.2% raw agreement and Cohen's $\kappa = 0.84$; the incident extension has 120 with 89.2% agreement and $\kappa = 0.76$.

Table S12: Independent transfer MR accuracy under two separately frozen retriever pools. Δ_{MR} is CHRONOPACKET minus Reader-Conditioned Selector with a paired-bootstrap 95% CI; reliability, source counts, and per-source results appear in Table S13.

Retriever	Method	MR ↑	Δ_{MR} [95% CI]
Contriever-v2	Long-Context Reader	59.1 ± 1.2	—
Contriever-v2	Reader-Conditioned Selector	62.4 ± 1.1	—
Contriever-v2	CHRONOPACKET	63.8 ± 1.0	1.4 [0.2, 2.8]
BGE-M3 (Chen et al., 2024)	Long-Context Reader	60.0 ± 1.1	—
BGE-M3 (Chen et al., 2024)	Reader-Conditioned Selector	63.0 ± 1.1	—
BGE-M3 (Chen et al., 2024)	CHRONOPACKET	64.5 ± 1.0	1.5 [0.4, 2.8]

The per-source results in Table S13 show consistent MR and answered-query stale-citation gains across release notes, policy bulletins, and knowledge-base updates. On the original set, CHRONOPACKET's paired MR advantage over Reader-Conditioned Selector is 1.4 [0.2, 2.8] points with Contriever-v2 and 1.5 [0.4, 2.8] with BGE-M3. Incident updates are harder in absolute terms, but the corresponding advantages increase to 2.9 and 2.8 points, with paired intervals excluding zero.

Table S13: Per-source and incident-update transfer results under frozen pools. Accuracy and packet reliability are separated into readable panels.

Retriever	Method	All/answered stale ↓	All-query contradiction ↓	Coverage
Contriever-v2	Long-Context Reader	14.6 ± 0.6 / 14.6±0.6	23.2±0.9	100.0
Contriever-v2	Reader-Conditioned Selector	10.8 ± 0.6 / 10.8±0.6	12.5±0.7	100.0
Contriever-v2	CHRONOPACKET	9.4 ± 0.5 / 9.9±0.5	7.8±0.6	94.4±0.6
BGE-M3	Long-Context Reader	14.1 ± 0.6 / 14.1±0.6	22.6±0.9	100.0
BGE-M3	Reader-Conditioned Selector	10.4 ± 0.6 / 10.4±0.6	11.9±0.7	100.0
BGE-M3	CHRONOPACKET	9.1 ± 0.5 / 9.7±0.5	7.4±0.6	94.1±0.6

Source family	Total	Standard	MR	Wrong-Newer	Missing-TS	Branch	Rollback
Software release notes	60	38	22	7	5	6	4
Policy bulletins	60	35	25	9	7	4	5
Knowledge-base status	60	35	25	8	6	5	6
Original-set total	180	108	72	24	18	15	15
Incident/status updates	240	132	108	36	24	28	20

Retriever / source	Method	MR ↑	WN ↑	Δ_{MR} [95% CI]
Contriever / release	Long-Context	60.7±1.3	54.0±1.6	—
	Reader-Conditioned	63.5±1.2	56.2±1.5	—
	CHRONOPACKET	65.1±1.1	58.0±1.4	1.6 [0.2, 3.2]
Contriever / policy	Long-Context	57.6±1.4	50.2±1.7	—
	Reader-Conditioned	61.3±1.3	54.3±1.6	—
	CHRONOPACKET	62.9±1.2	55.6±1.5	1.7 [0.1, 3.5]
Contriever / knowledge base	Long-Context	59.2±1.4	52.0±1.6	—
	Reader-Conditioned	62.4±1.3	55.4±1.5	—
	CHRONOPACKET	63.5±1.2	56.8±1.4	1.1 [0.1, 2.6]
Contriever / incidents	Long-Context	55.8±1.3	46.9±1.6	—
	Reader-Conditioned	60.2±1.2	54.4±1.5	—
	CHRONOPACKET	63.1±1.1	58.3±1.4	2.9 [1.4, 4.5]
BGE-M3 / incidents	Long-Context	57.0±1.2	47.8±1.5	—
	Reader-Conditioned	61.0±1.1	55.1±1.4	—
	CHRONOPACKET	63.8±1.0	58.9±1.3	2.8 [1.3, 4.4]

Retriever / source	Method	All/answered stale ↓	Contradiction ↓	Coverage
Contriever / release	Long-Context	14.1±0.7 / 14.1±0.7	23.0±0.9	100.0
	Reader-Conditioned	10.6±0.6 / 10.6±0.6	12.1±0.8	100.0
	CHRONOPACKET	9.0±0.5 / 9.5±0.6	7.4±0.7	94.7±0.7
Contriever / policy	Long-Context	15.2±0.8 / 15.2±0.8	24.1±1.0	100.0
	Reader-Conditioned	11.1±0.7 / 11.1±0.7	12.9±0.8	100.0
	CHRONOPACKET	9.8±0.6 / 10.4±0.7	8.2±0.7	93.9±0.8
Contriever / knowledge base	Long-Context	14.5±0.7 / 14.5±0.7	22.6±0.9	100.0
	Reader-Conditioned	10.7±0.6 / 10.7±0.6	12.4±0.8	100.0
	CHRONOPACKET	9.3±0.6 / 9.8±0.6	7.8±0.7	94.6±0.7
Contriever / incidents	Long-Context	15.8±0.7 / 15.8±0.7	25.2±0.9	100.0
	Reader-Conditioned	10.8±0.6 / 10.8±0.6	13.8±0.8	100.0
	CHRONOPACKET	8.9±0.6 / 9.6±0.6	8.1±0.7	92.4±0.7
BGE-M3 / incidents	Long-Context	15.2±0.7 / 15.2±0.7	24.5±0.9	100.0
	Reader-Conditioned	10.4±0.6 / 10.4±0.6	13.1±0.8	100.0
	CHRONOPACKET	8.5±0.5 / 9.2±0.6	7.7±0.6	92.7±0.7

Table S14 connects selective prediction to citation reliability and systems cost on the 8×A100 inference node used for Table 1. It reports both all-query and answered-query stale rates so abstention cannot receive hidden citation credit.

Table S14: Utility, overhead, and selective prediction under matched retrieval. Counts are seed means; rates are mean ± std over five seeds. Latency is measured over answered queries; selector-only overhead over all queries is reported in Table S4.

Method	Conflict q	Graph-path q	Reader-answer q	Mean packet	Abstained q
Long-Context Reader	—	—	600	20.0	0
CHRONOPACKET	208	208	562.8	2.8 ± 0.1	37.2

Method	p50 / p95 / p99 latency (ms)	Relative p50 cost
Long-Context Reader	284 / 412 / 503	1.00×
CHRONOPACKET	61 / 81 / 109	0.21×

Method	Answered n	Std. cov.	MR cov.	Ans. acc.	All stale	Ans. stale
Long-Context	600	100	100	72.6±0.6	13.2±0.5	13.2±0.5
Reader-Conditioned	600	100	100	74.7±0.6	9.1±0.4	9.1±0.4
CHRONOPACKET, no abst.	600	100	100	76.6±0.6	8.1±0.4	8.1±0.4
CHRONOPACKET	562.8	95.4±0.4	91.4±0.6	80.5±0.6	7.8±0.4	8.3±0.4

Method	Answered contradiction	Abstention precision
Long-Context	22.1±0.7	—
Reader-Conditioned	10.7±0.6	—
CHRONOPACKET, no abst.	7.0±0.5	—
CHRONOPACKET	6.7±0.5	82.1±1.0

Only 6.2% of queries are abstained, but those cases are wrong 82.1% of the time under forced answering. Conditioning on the 562.8 mean answered queries changes stale-citation incidence from 7.8% to 8.3%, still below every matched comparator. The selected operating point retains 95.4% Standard and 91.4% MR coverage, while answered-query median latency is 61 ms versus 284 ms for full-pool reading.

Table S15: Packet-formation sensitivity under matched retrieval. Every row uses the same fixed candidate pool and varies how interval tags, typed pruning, reordering, and abstention contribute to commitment.

Variant	MR Acc. ↑	Wrong-Newer Acc. ↑
CHRONOPACKET (full)	*66.4 ± 0.8*	*60.1 ± 0.9*
Fixed Contriever order + gate pruning (proxy)	65.8 ± 0.9	59.4 ± 1.0
Reader-only top-3 after claim extraction (proxy)	57.2 ± 1.0	48.6 ± 1.2

The first proxy preserves claim extraction, typed edges, pruning, and abstention while fixing Contriever-v2 order; the second delegates top-3 packet formation to the reader after interval tagging. Their separation shows that interval tags, typed pruning, and calibrated abstention jointly matter beyond positive reordering, while Table S1 assigns effects to single-factor interventions.