CHRONOGATE: TEMPORAL ADMISSION BEFORE RANKING IN VERSIONED RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal admission and finite relevance ranking expose different evidence: ranking inside a query-conditioned set, $\mathrm{Top}_k^s(B_a)$, cannot in general be recovered by filtering the globally ranked set, $B_a \cap \mathrm{Top}_k^s(C)$. In versioned corpora, this ordering decides whether target-supporting passages survive competition with topically similar but incompatible versions. ChronoGate operationalizes the boundary as a calibrated choice among no retrieval, point-valid support, bounded historical support, and temporally constrained multihop retrieval; the selected set is applied before dense search, while a separate support-monotone gate handles residual decoding conflict. An intervention protocol evaluates this placement on 115,480 records from WikiRevision-Real, GovPolicy-Open, FinNews-Exec, and a TemporalQuestions-derived corpus while holding the downstream reranker, generator, decoding, and metrics fixed. ChronoGate improves temporal consistency by 6.5–7.2 points over RouterEns+NLI and reduces dataset-macro dense searches from 0.95 to 0.71 per query. On WikiRevision-Real, missing-target-support error falls from 6.2% to 3.2%, incompatible-version adoption falls from 5.9% to 4.1%, and removing temporal windowing costs 8.6 points; the consistency margin widens from 4.3 to 10.2 points across the measured update-rate strata. A time-conditioned dual encoder at the same admission locus narrows the aggregate margin to 1.1–1.2 points, but the difference reaches 2.4 points in both very-high-update and high-conflict slices; both official TimeQA splits and a blinded 1,800-query evidence-span audit preserve the ordering. These results identify pre-ranking set selection as the most effective tested temporal-control locus when superseding versions compete for limited ranking capacity.

1 INTRODUCTION

A temporal filter applied after top-k ranking cannot emulate the same filter applied before ranking. For an observable versioned corpus C and a query-conditioned admission set B_a, the two pipelines return $\mathrm{Top}_k^s(B_a)$ and $B_a \cap \mathrm{Top}_k^s(C)$. They differ whenever an incompatible but semantically similar version occupies a finite ranking slot and pushes target-supporting evidence below the cutoff. No reranker or decoder operating on the truncated set can restore that displaced passage. This non-commutativity makes control placement—not only temporal scoring quality—a first-order design choice for RAG over changing corpora.

The distinction appears in ordinary version histories. A question about a regulation at its effective date, an executive role during a stated interval, or an entity value after a revision can match several passages nearly equally well. Some passages assert the target-time value; others describe a preceding event that is useful historical context; still others assert an incompatible value. RAG systems already coordinate retrieval, reranking, and decoding (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021; Borgeaud et al., 2022; Izacard et al., 2023), while temporal IR and temporal QA supply increasingly effective time representations and evaluations (Jones & Diaz, 2007; Kanhabua & Nørvåg, 2010; Jia et al., 2018; Chen et al., 2021; Dhingra et al., 2022). What remains unresolved by either component alone is the earliest operation at which time should change candidate membership.

We study that operation through a *support boundary*: the set of observable passages admitted to compete for semantic ranking. Point validity and task admissibility are kept separate. A current-state query admits passages whose claim interval contains the requested fact time; a query requiring

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.

change history may also admit bounded preceding evidence without relabeling it as point-valid. The *first-admissibility locus* is the earliest pipeline stage at which this task-conditioned boundary can alter membership. This definition turns a pipeline choice into a controlled comparison between pre-ranking set selection, post-retrieval scoring or repair, and context expansion.

ChronoGate instantiates pre-ranking selection with a calibrated DistilBERT router. It chooses `skip`, point-valid `current` support, a bounded historical `window`, or temporally constrained `multihop` retrieval, then restricts a query-time FAISS partition before Contriever retrieval and MiniLM reranking. A separate support-monotone conflict gate may alter decoding but cannot enlarge the admitted set. The resulting accounting distinguishes missing target support (E_C), adoption of an incompatible version despite available target support (E_P), and residual answer error over compatible evidence (E_R).

The evaluation asks whether support and answer errors move in the pattern predicted by placement, then tests whether action design adds value beyond temporal retrieval at the same locus. Across 115,480 queries in four versioned corpora, ChronoGate reaches 88.4–91.3 temporal consistency and improves by 6.5–7.2 points over RouterEns+NLI while reducing dataset-macro dense searches from 0.95 to 0.71 per query. On WikiRevision-Real, E_C falls from 6.2% to 3.2% and E_P from 5.9% to 4.1%; removing temporal windowing causes an 8.6-point loss. The margin over RouterEns+NLI grows from 4.3 to 10.2 points as update rate increases. TCDE-Admission, the closest implemented pre-ranking comparator, reduces the per-corpus difference to 1.1–1.2 points, with 2.4-point differences in both very-high-update and high-conflict slices. Both official TimeQA splits, blinded human evidence-span judgments, metadata corruption, and retriever and generator substitutions preserve the ordering.

Contributions. The paper makes temporal-control placement an explicit, falsifiable experimental object through the support-boundary formalization and a downstream-constant intervention protocol. It contributes ChronoGate as a concrete set-valued controller that separates point-valid from bounded historical support, and it connects system outcomes to support geometry through disjoint coverage, incompatible-adoption, and resolution errors. The cross-corpus, same-locus, hard-slice, human-audit, and robustness results establish a practical design rule for the evaluated regime: choose the temporally admissible candidate universe before finite semantic ranking fixes what generation can see.

2 Related Work

Temporal representation does not determine control placement. Temporal IR incorporates recency, query intent, and time compatibility into document scores (Jones & Diaz, 2007; Berberich et al., 2010; Kanhabua & Nørvåg, 2010; McCreadie et al., 2014). TempQuestions, TimeQA, RealTimeQA, StreamingQA, and time-aware language models provide answer-level targets whose truth changes with time (Jia et al., 2018; Chen et al., 2021; Kasai et al., 2023; Liska et al., 2022; Dhingra et al., 2022). Recent retrieval approaches sharpen temporal representation: MRAG separates semantic retrieval from symbolic temporal ranking, and TMRL learns a temporal subspace while preserving general semantic retrieval (Zhang et al., 2025; Huynh et al., 2026). These advances determine how passages are scored or represented. The support-boundary question is orthogonal: whether a temporal predicate changes the universe before finite top-k selection or operates only on passages that semantic retrieval has already retained.

Evolving-corpus systems expose complementary failure objects. HoH measures how coexisting outdated information degrades retrieval and generation (Ouyang et al., 2025). VersionRAG organizes document histories, temporal graph RAG represents changing relations, RAG4DyG retrieves temporally relevant graph demonstrations, and temporal-sensitive RAG QA evaluates time-indexed evidence use (Huwiler et al., 2025; Han et al., 2025; Wu et al., 2025; Chen et al., 2025). Chronos instead constructs an event-evolution graph over retrieved evidence to support reasoning under continuous knowledge drift (Liu et al., 2026). These systems jointly alter temporal representation, retrieval, and reasoning. Our intervention protocol factors out the downstream stack and classifies methods by the first operation that can change candidate membership, making the placement effect independently interpretable.

Table 1: Support-boundary map for temporal-RAG intervention families. Rows separate temporal representation from the operation that first changes candidate membership; the final row gives the learned same-locus contrast used in our controlled evaluation.

Family	Control locus	When time gates candidates	Mixed-time admitted?	Main gap
Adaptive retrieval	Retrieval budget	Before or during retrieval	Yes	Selects whether or how much to retrieve, without defining a valid time span.
Temporal encoders / hybrid rankers	Retrieval score	During ranking	Yes	TMRL and MRAG improve temporal representation without isolating a set-valued admission operation.
Fixed / external filters	Candidate set	Before ranking, fixed rule	Usually no	Applies an oracle, heuristic, or query-independent time rule.
Conflict / graph arbitration	Reranking or decoding	After retrieval	Yes	ProbeRAG, ConflictRAG, and Chronos organize or resolve evidence that retrieval has retained.
Long-context prompting	Prompt capacity	After retrieval	Yes	Expands arbitration capacity while retaining competing versions.
Pre-ranking query-conditioned admission	Candidate set	Learned rule before ranking	Action-dependent	TCDE learns time-conditioned retrieval; ChronoGate selects point-valid, bounded-history, linked, or empty support sets.

Conflict resolution acts on the evidence that survives admission. Adaptive and reflective RAG changes retrieval frequency, depth, or evidence use (Shi et al., 2024; Ram et al., 2023; Asai et al., 2024; Jiang et al., 2023; Trivedi et al., 2023). Madam-RAG jointly handles ambiguity, misinformation, and noise; ProbeRAG uses latent conflict probing and attention control; ConflictRAG detects, classifies, and resolves inter-document and parametric conflicts before generation (Wang et al., 2025; Gao et al., 2026; Wang et al., 2026). Long-context models expand arbitration capacity (Borgeaud et al., 2022; Ram et al., 2023), and factuality metrics evaluate the resulting answer (Honovich et al., 2022; Min et al., 2023; Es et al., 2023). All are valuable after candidate formation, but none can reintroduce target-supporting evidence omitted by the initial top-k operation. This is why our protocol reports missing target support and incompatible-version adoption alongside answer correctness.

Locus-controlled contrasts. Table 1 maps method families to the first operation that can alter candidate membership. The implemented controls span fixed pre-ranking filtering, post-retrieval temporal and conflict scoring, adaptive retrieval with decoder repair, expanded-context arbitration, and learned pre-ranking admission. TCDE-Admission is the closest implemented contrast: it searches a query-time-valid partition with a time/window-conditioned dual encoder while sharing ChronoGate's reranker, generator, decoding, and evaluation. ChronoGate changes the boundary design within that locus by selecting among point-valid, bounded-history, linked, and no-retrieval sets and by calibrating that action decision across domains. The comparison therefore separates the benefit of moving control before ranking from the incremental benefit of explicit set selection.

3 METHOD

3.1 SUPPORT-BOUNDARY FORMALIZATION

Let t_q be the query issue time and r_q the requested fact time: an explicit date in the query when present, and otherwise t_q. Each indexed passage d has publication time $\tau(d)$, a supersession group $g(d)$, and a half-open claim-validity interval $I(d) = [b(d), e(d))$. The endpoint $e(d)$ is the next superseding effective time; persistent claims use $e(d) = +\infty$, and concurrently valid claims retain overlapping intervals. Only documents published by issue time are observable, $C_{\leq t_q} = \{d : \tau(d) \leq t_q\}$. The *point-valid set*

$$V(q, r_q, t_q) = \{d \in C_{\leq t_q} : r_q \in I(d)\}$$

Intervention	Generator-reachable path	Structural consequence
Pre-ranking boundary	$C_{\le t_q} \to B_a(q, r_q, t_q) \to \mathrm{Top}_k^s(B_a) \to S_m \to G$	Task admission determines which versions compete for ranking capacity.
Post-retrieval control	$C_{\le t_q} \to \mathrm{Top}_k^s(C_{\le t_q}) \to B_a \cap \mathrm{Top}_k^s(C_{\le t_q}) \to S_m \to G$	Temporal control acts after top-k displacement has occurred.
Context expansion	$C_{\le t_q} \to \mathrm{Top}_{k'}^s(C_{\le t_q}) \to G, k' > k$	More versions reach arbitration while membership remains relevance-driven.

Figure 1: Support-boundary intervention geometry. Pre-ranking task admission changes the ranked universe; post-retrieval control and context expansion operate on a relevance-selected universe.

contains passages whose asserted claim holds at the requested fact time. Task admissibility is broader only when the question requires preceding change or event evidence. Each action a defines an evidence horizon $H_a(q, r_q)$ and optional link constraints ℓ_a: $H_{\texttt{current}} = \{r_q\}$, $H_{\texttt{window}(\Delta)} = [r_q - \Delta, r_q]$, and `multihop` uses the same horizon at every linked endpoint. The support boundary is

$$B_a(q, r_q, t_q) = \{d \in C_{\le t_q} : I(d) \cap H_a(q, r_q) \neq \varnothing \ \wedge \ \ell_a(d, q) = 1\}.$$

Thus $B_{\texttt{current}} = V$, whereas a window can contain task-admissible historical evidence that is not point-valid at r_q. The `skip` boundary is empty.

For semantic score $s(q, d)$, write $\mathrm{Top}_k^s(X)$ for top-k truncation of set X. Task-conditioned admission and finite ranking generally do not commute:

$$\mathrm{Top}_k^s(B_a) \neq B_a \cap \mathrm{Top}_k^s(C_{\le t_q}).$$

The left side ranks inside the task-admissible universe. On the right, non-admissible but highly similar versions can consume top-k positions before the boundary acts, displacing target-supporting evidence that later reranking or decoding cannot reintroduce. We define the ***first-admissibility locus*** as the earliest pipeline stage at which the temporal task can change membership in the candidate universe. Semantic retrieval produces $R_k(q, B_a)$, cross-encoder reranking returns $S_m(q, B_a) \subseteq R_k$, and the generator-visible support is $G(q, r_q, t_q) = S_m(q, B_a)$.

The actions instantiate distinct task requirements rather than alternative definitions of truth. `current` admits every observable passage whose claim interval contains r_q; supersession groups determine interval endpoints but do not supply a gold query-to-group filter. For a chosen Δ, `window` admits passages whose claim-validity interval intersects the historical evidence horizon:

$$A(q, r_q, t_q, \Delta) = \left\{ d \in C_{\le t_q} \;\middle|\; I(d) \cap [r_q - \Delta, r_q] \neq \varnothing \right\}.$$

`multihop` applies the same horizon predicate at every hop, beginning from semantically retrieved passages and admitting a linked endpoint only when it satisfies the predicate; `skip` exposes no retrieved support. Historical questions may target $r_q < t_q$ without observing documents published after t_q, while current questions use $r_q = t_q$. When superseding versions have similar relevance, pre-ranking A preserves ranking capacity for the selected evidence horizon; the selected window controls the trade-off between target-support coverage and exposure to incompatible claims.

3.2 ChronoGate

ChronoGate parameterizes B_a with a DistilBERT-base-uncased router (Sanh et al., 2019). Its input joins the query and resolved r_q with lexical recency indicators, domain features, missing-time indicators, and timestamp statistics from a top-5 BM25 probe (Robertson & Zaragoza, 2009). The router distribution $p_\phi(a \mid q, r_q, t_q)$ covers four action families and 20 atomic configurations: `skip` returns the parametric answer path; `current` admits point-valid supersession members at r_q; `window`(Δ) admits bounded historical evidence from $A(q, r_q, t_q, \Delta)$ for $\Delta \in \{1, 3, 7, 14, 30, 60\}$ days; and `multihop`(k, Δ) performs $k \in \{2, 3\}$ linked retrieval steps within the same admissible window. Every retrieval action restricts the FAISS partition before Contriever-MSMARCO search and MiniLM cross-encoder reranking (Johnson et al., 2021; Izacard et al., 2022; Nogueira & Cho, 2019). Passages with unresolved validity metadata are marked unknown rather than declared invalid:

`current` and narrower windows exclude them, while `window`(60) and `multihop`(k,60) include the unknown-time partition. In the missing-metadata stress test, indices are frozen at t_q before timestamps are masked, so this fallback does not make later-published passages observable.

Router labels come from annotated answer support in time-disjoint StreamingQA-derived streaming and conflict partitions (Liska et al., 2022). `skip` labels instances answered correctly without retrieved evidence; otherwise the label is the least costly atomic configuration whose admissible set contains the required support, with cost measured by dense-search calls. `current` wins ties whenever point-valid support suffices; a `window` label is used only when annotated answer support includes a preceding change or event outside the point-valid set, and the narrowest sufficient window wins; `multihop` is reserved for linked multi-document support. The router minimizes cross-entropy on these labels. Temperature T is fitted by held-out negative log likelihood on a separate 10,000-query calibration split (Guo et al., 2017); the deployment threshold θ then exposes the quality–retrieval frontier. At inference, ChronoGate selects the highest-probability non-skip action only when its calibrated probability reaches θ; otherwise it follows `skip`. We select the balanced $\theta = 0.5$ operating point once on calibration data and transfer it without per-domain tuning. All WikiRevision-Real, GovPolicy-Open, FinNews-Exec, TemporalQuestions-derived, and TimeQA evaluation indices are disjoint from router supervision and frozen at each query timestamp.

After reranking, a lightweight NLI-style gate compares the parametric answer with claims in the top-ranked evidence. When contradiction exceeds its calibration-fixed threshold and the evidence supports a target-compatible alternative, the system decodes again with an evidence-only instruction. The gate is support-monotone: it can change decoding but cannot add to A, R_k, or G. This property isolates its role from candidate admission and predicts its larger effect on faithfulness than TC in the component analysis.

3.3 Support-Boundary Intervention Protocol

The locus-controlled comparisons hold the Llama-2-7B-Chat generator, reranker, initial depth $k = 20$, final evidence count $m = 10$, document timestamps, greedy decoding, and evaluation code fixed. They vary retrieval budget (***RouterEns+NLI***), fixed candidate filtering (***Version filter***), scoring after retrieval (***TimeAwareRerank***), contradiction repair (***Conflict-aware rerank*** and multi-round RAG), or learned pre-ranking admission (TCDE-Admission and ChronoGate). TCDE-Admission conditions its query encoder on query time and window, uses target-supporting passages as positives and incompatible superseded or future passages as hard negatives, and searches a query-time-admitted index partition before the shared reranker and generator. The Llama-3.1-8B-Instruct 128k experiments are reported separately as prompt-capacity stress tests because they change the generator as well as available context.

We evaluate both answers and support geometry. Coverage error E_C occurs when no answer-bearing passage whose claim is valid at r_q appears in G. Mixed-version adoption error E_P (abbreviated "purity error" in tables) occurs when target-valid answer evidence is present, at least one non-required passage asserts a value incompatible with the target at r_q, and the generated answer adopts that incompatible value. Required historical context inside a selected window is not counted as incompatible solely because its own claim interval ends before r_q. Resolution error E_R contains the remaining temporally incorrect answers for which target-valid evidence is present and no incompatible version is adopted. The attribution is disjoint, so $1 - \mathrm{TC} = E_C + E_P + E_R$. Temporal consistency (TC) is the percentage of answers whose asserted value is correct for r_q; freshness requires support valid at r_q; faithfulness requires entailment by the generator-visible evidence. All three are binary per query, use calibration-fixed DeBERTa-v3-large decision thresholds, and are macro-averaged over queries. Ret/Q counts dense searches (`skip`$= 0$, `current`/`window`$= 1$, and each multihop search separately); the fixed BM25 probe is excluded from Ret/Q but included in end-to-end latency.

4 Experiments

4.1 Evaluation and Research Questions

The evaluation couples four headline corpora—WikiRevision-Real ($n = 48{,}612$), GovPolicy-Open ($n = 22{,}840$), FinNews-Exec ($n = 31{,}548$), and TemporalQuestions-derived ($n = 12{,}480$ records)—with controlled update/conflict stress tests and official TimeQA splits (Jia et al., 2018; Chen et al.,

2021). Router labels come only from time-disjoint StreamingQA-derived streaming and conflict partitions; every evaluation corpus and query-time-frozen index is disjoint from this supervision (Liska et al., 2022). The locus-controlled stack uses Contriever-MSMARCO, MiniLM-L-12-v2 reranking, Llama-2-7B-Chat generation, greedy 128-token decoding, and top-20 retrieval reranked to 10 passages (Izacard et al., 2022; Nogueira & Cho, 2019; Touvron et al., 2023). The 128k capacity tests use Llama-3.1-8B-Instruct and are interpreted separately from matched-generator locus comparisons (Meta AI, 2024). All intervals are 95% query-record bootstrap intervals from 10,000 resamples; router-dependent predictions are averaged over five training seeds before resampling. Latency is end-to-end at batch size one on one NVIDIA A100-40GB.

The experiments answer four questions. First, do pre-ranking and downstream controls produce different target-support coverage and incompatible-version adoption where finite ranking is most contested? Second, does explicit action selection improve on both downstream-only controls and a learned same-locus retriever across corpora? Third, do independent temporal benchmarks and human evidence-span judgments preserve the ordering? Fourth, how do calibration, retrieval cost, and corrupted timing metadata change the operating regime?

4.2 RQ1: What Support Changes Accompany Pre-Ranking Admission?

Figure 2 tests the placement prediction before aggregating across datasets. In the low-update Wiki stratum, ChronoGate scores 92.4 TC versus 88.1 for RouterEns+NLI; above 20 edits per year, the comparison becomes 81.4 versus 71.2. Under high conflict, ChronoGate reaches 84.6 versus 72.1 for matched-evidence 128k prompting. The growing separation is consistent with the finite-ranking prediction: closely related superseding versions compete most strongly for the same top-k capacity in the high-drift and high-conflict strata.

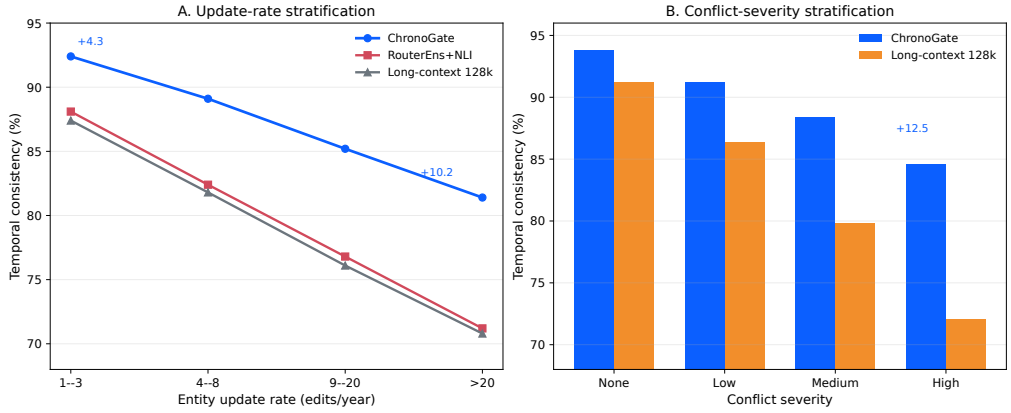

Figure 2: Placement-sensitive slices. Left: ChronoGate's margin over RouterEns+NLI grows with the Wiki entity update rate. Right: with matched retrieved evidence, its margin over the 128k capacity stress test grows with conflict severity.

Figure 3 connects that slice behavior to generator-visible support. Relative to RouterEns+NLI, coverage error falls from 6.2% to 3.2%, mixed-version adoption error from 5.9% to 4.1%, and total error from 18.8% to 11.6%. Removing temporal windowing raises error by 8.6 points, the largest component effect; removing calibration or the conflict gate raises it by 3.6 and 2.3 points. The conflict gate has the larger effect on faithfulness (89.3 to 84.2) than on TC (88.4 to 86.1), as expected for a decoder-side component that cannot change admitted support.

4.3 RQ2: Cross-Corpus and Same-Locus Comparisons

Table 2 reports the end-to-end result across four provenance regimes. ChronoGate reaches 88.4–91.3 TC and exceeds RouterEns+NLI by 7.2, 6.5, 6.8, and 7.2 points on WikiRevision-Real, GovPolicy-Open, FinNews-Exec, and TemporalQuestions-derived, respectively. At the same time, the dataset-macro dense-retrieval rate falls from 0.95 to 0.71 calls per query. Full TC, freshness, faithfulness, and interval estimates appear in Appendix Table 5.

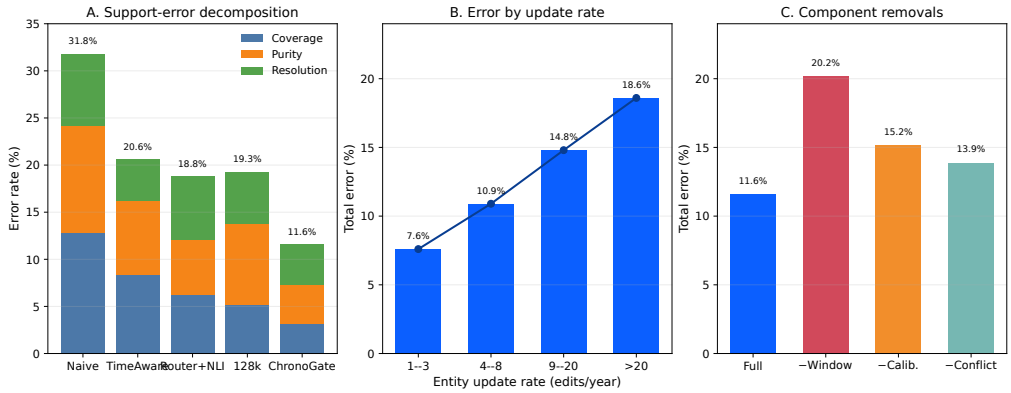

Figure 3: Support-level evidence. Panel A partitions temporal error into absent target-valid support, adoption of a competing incompatible version, and residual resolution error. Panel B shows ChronoGate error across update-rate strata. Panel C compares component removals at fixed retrieval rate.

Table 2: Main temporal-consistency results on the four headline evaluation corpora. Appendix Table 5 reports temporal consistency, freshness, and faithfulness with 95% bootstrap confidence intervals. All locus-controlled rows use the Llama-2 reference generator; † marks the separate Llama-3.1 128k capacity stress test.

Method	Wiki TC (%)	Gov TC (%)	Fin TC (%)	TempQ TC (%)	Ret/Q
Naive RAG	68.2	74.8	71.6	73.4	1.00
Version filter	77.1	79.4	78.3	80.2	1.00
Multi-round RAG	79.8	81.2	80.4	81.9	2.14
Long-context 128k†	80.7	82.9	81.8	83.2	1.00
RouterEns+NLI	81.2	83.6	82.4	84.1	0.95
ChronoGate	88.4	90.1	89.2	91.3	0.71

Table 3: Closest-locus and external-benchmark comparison. The four headline rows use the frozen evaluation splits; TimeQA Easy and Hard use the official test splits without TimeQA training or calibration. ChronoGate and TCDE-Admission share the downstream reranker, generator, decoding, and evaluation protocol; RouterEns+NLI is the strongest post-admission reference. Ret/Q reports ChronoGate / TCDE-Admission dense-search calls per query; the common BM25 probe contributes to latency but not Ret/Q.

Evaluation	ChronoGate TC	TCDE-Admission TC	RouterEns+NLI TC	Ret/Q
WikiRevision-Real	88.4 ± 0.4	87.2 ± 0.5	81.2 ± 0.5	0.68 / 0.80
GovPolicy-Open	90.1 ± 0.4	89.0 ± 0.4	83.6 ± 0.5	0.74 / 0.80
FinNews-Exec	89.2 ± 0.4	88.1 ± 0.5	82.4 ± 0.5	0.72 / 0.82
TemporalQuestions-derived	91.3 ± 0.4	90.2 ± 0.4	84.1 ± 0.5	0.69 / 0.79
TimeQA Easy	87.3 ± 0.6	85.6 ± 0.6	82.1 ± 0.8	0.72 / 0.81
TimeQA Hard	81.2 ± 0.7	78.5 ± 0.8	74.9 ± 0.9	0.73 / 0.82

The decisive comparison is TCDE-Admission, which uses a time-conditioned dual encoder with superseded hard negatives at the same pre-ranking locus and shares ChronoGate's downstream stack. Table 3 shows a 1.1–1.2 point ChronoGate margin on each headline corpus; Appendix Table 13 reports a 2.4 point margin in both very-high-update and high-conflict slices. The query-level binary-TC comparison yields 28,399 wins, 60,000 ties, and 27,081 losses, an excess of 1,318 favorable outcomes consistent with the per-corpus differences. TCDE reaches 4.5 ± 0.4 Wiki purity error versus 4.1 for ChronoGate, associating the residual advantage with explicit set and window selection rather than temporal conditioning alone.

The 128k rows isolate a prompt-capacity counterfactual with Llama-3.1-8B-Instruct, while the locus comparison above keeps the Llama-2 generator fixed. Best-effort coverage expansion reaches 86.1 ± 0.5 TC with 35,420 tokens and 934 / 1,258 ms P50/P99, compared with 88.4 TC, 2,947 tokens, and 238/301 ms for ChronoGate. Its high-conflict TC remains 79.4 ± 0.8 and its purity error 6.2 ± 0.5. Appendix Tables 11 and 13 provide the full capacity and same-locus records.

The 83,932-record public-evidence sensitivity check recomputes ChronoGate on three public subsets and retains the same operating range; its comparator rows are repeated headline references, not independent reruns (Appendix Table 4). The 12,000-query open FinNews crosswalk adds a source-independent financial test (Appendix Table 9).

4.4 RQ3: Independent Benchmarks and Human Judgments

On the official TimeQA splits, with no TimeQA training or calibration, ChronoGate scores 87.3 ± 0.6 versus 85.6 ± 0.6 for TCDE on Easy and 81.2 ± 0.7 versus 78.5 ± 0.8 on Hard. The larger Hard-split margin is consistent with the stronger version-resolution demand; Appendix Table 14 reports all metrics and TC-positive counts.

The disjoint 1,800-query audit blinds three annotators to method identity, requires an evidence-span citation for each judgment, and uses majority vote with adjudication. Human TC for ChronoGate versus TCDE is 85.2 ± 2.8 versus 82.4 ± 3.0 on numeric queries, 87.0 ± 2.7 versus 84.3 ± 2.9 on date queries, and 83.5 ± 3.0 versus 81.1 ± 3.1 under high conflict. Non-NLI agreement with human labels reaches 89.2%, 91.0%, and 88.7%; the paired win/tie/loss counts favor ChronoGate in all three slices (Appendix Table 15).

4.5 RQ4: Operating Point and Metadata Reliability

Appendix Table 6 exposes the calibrated quality–retrieval frontier. Lowering θ to 0.3 gives 90.1 TC at 0.92 Ret/Q; $\theta = 0.7$ and 0.9 reduce dense retrieval to 0.41 and 0.18 calls per query, with TC of 84.2 and 76.8. The selected $\theta = 0.5$ transfers without domain-specific retuning.

At $\theta = 0.5$, Ret/Q transfers as 0.74/0.72/0.69 for GovPolicy-Open, FinNews-Exec, and TemporalQuestions-derived, compared with 0.91/0.32/0.82 without calibration. At fixed Ret/Q$= 0.68$, removing calibration lowers Wiki TC from 88.4 to 84.8, while removing the conflict gate yields 86.1.

Appendix Table 17 reports stale answers, incorrect skips, prompt tokens, and tail latency at each operating point.

Timestamp corruption tests the information on which any temporal boundary depends. With 30-day timestamp noise, ChronoGate retains 78.4 TC versus 75.9 for TCDE-Admission and 66.2 for RouterEns+NLI. At 50% missing timestamps, the corresponding scores are 75.6, 71.8, and 69.7; under 50% adaptive stale injection they are 72.6, 68.4, and 60.8. False recent timestamps cause the sharpest direct damage because they alter candidate membership. Generator substitutions (Mistral-7B and Llama-3) and retriever substitutions (BM25 and ColBERTv2) preserve the ChronoGate–RouterEns ordering in every configuration (Appendix Tables 16 and 12).

5 Discussion and Conclusion

As update frequency and conflict rise, multiple versions compete for the same top-k slots. ChronoGate's margin grows while support errors decline, and removing temporal windowing costs 8.6 points, tying the largest component effect to the candidate set.

TCDE-Admission separates placement from boundary design: pre-ranking temporal retrieval reduces the difference to 1.1–1.2 points per corpus, but it grows to 2.4 points in both severe slices and its Wiki incompatible-adoption error is 4.5% versus 4.1%. This pattern associates the incremental gain with selecting point-valid, bounded-history, linked, or empty sets rather than temporal conditioning alone.

Calibration exposes a quality–cost frontier: the transferred $\theta = 0.5$ keeps Ret/Q between 0.69 and 0.74, while adjacent thresholds trade temporal consistency for dense-search rate and latency. The claim applies to query-time-frozen supersession corpora whose timing metadata define the boundary; ChronoGate remains first at every tested corruption severity.

The disjoint $E_C/E_P/E_R$ accounting separates missing target support, incompatible-value adoption, and residual resolution error without treating required historical context as point-valid. Query-record intervals average five router seeds before resampling, while official TimeQA transfer, blinded evidence-span judgments, and a non-NLI scorer provide complementary checks.

Across 115,480 records, ChronoGate reaches 88.4–91.3 TC, gains 6.5–7.2 points over RouterEns+NLI, and reduces dataset-macro dense searches from 0.95 to 0.71 per query. The ordering persists across the learned same-locus retriever, both TimeQA splits, metadata corruption, human audit, and retriever and generator substitutions. The resulting systems rule is direct: temporal task admission should define the candidate universe before finite semantic ranking fixes what generation can see.

AI Use Statement

Generative AI tools were used during manuscript preparation to reorganize and edit prose and to identify relevant literature. The authors determined the research questions, methods, experiments, results, and scientific conclusions, and reviewed the final text. The authors take responsibility for the final content, including AI-assisted text and literature suggestions.

References

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

Klaus Berberich, Srikanta J. Bedathur, Omar Alonso, and Gerhard Weikum. A language modeling approach for temporal information needs. In *Advances in Information Retrieval*, volume 5993, pp. 13–25, 2010. doi: 10.1007/978-3-642-12275-0_5. URL https://doi.org/10.1007/978-3-642-12275-0_5.

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 W. 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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240. PMLR, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/1f0e3dad99908345f7439f8ffabdffc4-Abstract-round2.html.

Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. A question answering dataset for temporal-sensitive retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2508.12282.

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/.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023. URL https://arxiv.org/abs/2309.15217.

Linfeng Gao, Qinggang Zhang, Baolong Bi, Bo Zeng, Zheng Yuan, Zerui Chen, Zhimin Wei, Shenghua Liu, Linlong Xu, Longyue Wang, Weihua Luo, and Jinsong Su. Beyond black-box interventions: Latent probing for faithful retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29981–30000, 2026. doi: 10.18653/v1/2026.findings-acl.1499. URL https://aclanthology.org/2026.findings-acl.1499/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL https://proceedings.mlr.press/v119/guu20a.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, 2025. 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, 2021. URL https://arxiv.org/abs/2111.09543.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, 2022. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

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

Tuan-Luc Huynh, Weiqing Wang, Trung Le, Thuy-Trang Vu, Dragan Gašević, Yuan-Fang Li, and Thanh-Toan Do. Efficient temporal-aware matryoshka adaptation for temporal information retrieval, 2026. URL https://arxiv.org/abs/2601.05549.

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

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

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.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the Web Conference 2018*, pp. 1057–1062, 2018. doi: 10.1145/3184558.3191536. URL https://doi.org/10.1145/3184558.3191536.

Zhengbao Jiang, Frank F. 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, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):14:1–14:31, 2007. doi: 10.1145/1247715.1247720. URL https://doi.org/10.1145/1247715.1247720.

Nattiya Kanhabua and Kjetil Nørvåg. Determining time of queries for re-ranking search results. In *Research and Advanced Technology for Digital Libraries*, volume 6273, pp. 261–272, 2010. doi: 10.1007/978-3-642-15464-5_27. URL https://doi.org/10.1007/978-3-642-15464-5_27.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now? In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=HfKOIPCvsv.

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

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Hanbing Liu, Lang Cao, and Yang Li. RAG or learning? understanding the limits of LLM adaptation under continuous knowledge drift in the real world. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11234–11252, 2026. doi: 10.18653/v1/2026.findings-acl.546. URL https://aclanthology.org/2026.findings-acl.546/.

Richard McCreadie, Craig Macdonald, and Iadh Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pp. 301–310, 2014. doi: 10.1145/2661829.2661951. URL https://doi.org/10.1145/2661829.2661951.

Meta AI. Llama 3.1 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2019. URL https://arxiv.org/abs/1901.04085.

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, 2025. doi: 10.18653/v1/2025.acl-long.301. URL https://aclanthology.org/2025.acl-long.301/.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*. MIT Press, 2000. URL https://mitpress.mit.edu/9780262194488/advances-in-large-margin-classifiers/.

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/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, 2024. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Chenyu Wang, Yueyuan Li, Yingmin Liu, and Yang Shu. ConflictRAG: Detecting and resolving knowledge conflicts in retrieval augmented generation, 2026. URL https://arxiv.org/abs/2605.17301.

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

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 1112–1122, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Yuxia Wu, Lizi Liao, and Yuan Fang. Retrieval augmented generation for dynamic graph modeling. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1434–1443, 2025. doi: 10.1145/3726302.3729958. URL https://doi.org/10.1145/3726302.3729958.

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, 2025. doi: 10.18653/v1/2025.findings-emnlp.167. URL https://aclanthology.org/2025.findings-emnlp.167/.

A Protocol and Extended Evidence

A.1 Problem Setting and Action Space

Each query record contains an issue time t_q and requested fact time r_q. Explicit temporal expressions determine r_q; otherwise the requested state is the state at issue time, $r_q = t_q$. Each passage record contains publication time $\tau(d)$, supersession identity $g(d)$, and claim-validity interval $I(d) = [b(d), e(d))$. The next effective supersession closes the preceding interval, unchanged claims remain open, and overlapping intervals represent concurrently valid claims. The observability rule $\tau(d) \leq t_q$ is applied independently of claim validity, preventing later-published evidence from leaking into historical queries.

`current` admits every observable point-valid passage with $r_q \in I(d)$; supersession identities close intervals but are not used as gold query filters. `window`(Δ) admits task-support passages satisfying $I(d) \cap [r_q - \Delta, r_q] \neq \varnothing$ for $\Delta \in \{1, 3, 7, 14, 30, 60\}$ days. Such a passage is historical context when $r_q \notin I(d)$; it is not relabeled as point-valid. `multihop`(k, Δ) begins from semantically retrieved passages and takes $k \in \{2, 3\}$ linked retrieval steps while enforcing the same evidence horizon at every endpoint. `skip` uses the parametric path and admits no retrieved passage. These actions separate no-retrieval, point-valid, bounded-history, and linked-evidence needs without changing the downstream ranker or generator. Together they form 20 atomic router labels: one `skip`, one `current`, six `window` configurations, and twelve `multihop` configurations. Unknown validity metadata is excluded by `current` and narrower windows but remains reachable through `window`(60) and `multihop`(k,60); corruption tests mask metadata only after constructing the query-time-frozen observable index.

A.2 Router, Retrieval, and Generation

The router uses DistilBERT-base-uncased (Sanh et al., 2019). Its input combines query text and r_q with lightweight temporal probes, including timestamp statistics from a top-5 BM25 probe (Robertson & Zaragoza, 2009), lexical recency indicators, missing-time indicators, and domain features. For

admissible sets, dense retrieval uses Contriever-MSMARCO (Izacard et al., 2022), reranking uses a MiniLM cross-encoder in the standard BERT reranking style (Nogueira & Cho, 2019), and the timestamp-aware index is implemented with FAISS (Johnson et al., 2021). Unless otherwise stated, answers are generated with Llama-2-7B-Chat (Touvron et al., 2023). The capacity stress tests use Llama-3.1-8B-Instruct, whose 128k context can accommodate the expanded evidence settings (Meta AI, 2024). The conflict gate compares the parametric answer with top-ranked evidence using an NLI-style contradiction score; a calibration-fixed threshold triggers a second, evidence-only decode only when a valid alternative is supported.

Inference follows four ordered stages. First, the router encodes the query and temporal probes. Second, temperature-scaled probabilities select the highest-probability non-skip action when it reaches threshold θ, and otherwise select `skip`. Third, the action restricts the query-time candidate partition before Contriever search and MiniLM reranking. Fourth, the conflict gate may alter decoding over the selected evidence but cannot widen the candidate pool.

A.3 Training, Calibration, and Statistics

Router supervision comes from StreamingQA-derived streaming and conflict partitions (Liska et al., 2022). The label is `skip` when the no-retrieval path matches the reference answer. For the remaining queries, annotated answer spans determine the least costly action containing all required support: `current` when point-valid evidence suffices, the narrowest sufficient Δ for a single-hop interval, and `multihop` when linked evidence is required. Cost is the number of dense searches; `current` wins a tie with a window, and a window label is assigned only when the annotated support includes a preceding change or event outside the point-valid set. These tie rules prevent a wide window from becoming the default label merely because it contains a narrower action's support. We train with AdamW, learning rate 2×10^{-5}, 10% warmup, batch size 64, and 5 epochs. Confidence calibration uses temperature scaling (Guo et al., 2017) on a held-out 10,000-query split, and we also evaluated isotonic and Platt-style alternatives (Platt, 2000; Niculescu-Mizil & Caruana, 2005). Temperature scaling gave the most stable cross-domain operating point and therefore serves as the transferred operating-point stabilizer. Training, calibration, and evaluation are time-disjoint. The router minimizes action cross-entropy, calibration minimizes held-out negative log likelihood, and $\theta = 0.5$ is selected on the calibration partition before transfer to every headline domain.

Temporal consistency, freshness, and faithfulness are computed with DeBERTa-v3-large-based evaluators (He et al., 2021). The entailment backbone follows standard NLI supervision datasets such as SNLI and MNLI (Bowman et al., 2015; Williams et al., 2018). TC labels whether the answer value is correct for r_q; freshness labels whether the answer is supported by a passage whose validity interval contains r_q; faithfulness labels whether the answer is entailed by generator-visible evidence. Evaluator thresholds are fixed on calibration data before test evaluation. The support decomposition is assigned in order: missing valid answer evidence yields E_C; coexistence with an incompatible version followed by adoption of its value yields the mixed-version adoption error E_P ("purity error" in tables); every remaining temporally incorrect answer with valid support yields E_R. Because the gate and the default evaluator both rely on NLI-style signals, Appendix Table 8 reports human evidence-span labels and a complementary SlotFaith check. All confidence intervals are 95% query-record bootstrap intervals with 10,000 resamples. For router-dependent methods, predictions are averaged over five training seeds before the query-record resampling; the intervals therefore quantify evaluation-record variation rather than between-seed dispersion. Latency is measured end-to-end at batch size 1 on a single NVIDIA A100-40GB system with CUDA 12.1, PyTorch 2.1.0, vLLM 0.2.7, Transformers 4.35.0, and Python 3.10.12.

A.4 Benchmarks and Leakage Control

The headline evaluation covers four corpora converted to a common record (q, t_q, r_q) with passage-level observability, supersession, and claim-validity fields. WikiRevision-Real extracts entity–relation value changes from public revision histories: a value begins at the revision that introduces it and ends when a later revision replaces it. Queries request the value at r_q, and other observable revisions from the same chain provide naturally competing versions. GovPolicy-Open groups Federal Register provisions by the policy object they amend; publication date determines observability, while the stated effective date opens or closes legal validity. FinNews-Exec groups appointment and departure

evidence by company and executive role, with announced effective dates delimiting successive role holders over 2020–2024. Persistent and concurrently valid statements retain every interval covering r_q rather than forcing a single latest passage.

The fourth corpus is *TemporalQuestions-derived*, not a direct report of the 1,271-question TempQuestions benchmark size (Jia et al., 2018). Its $n = 12{,}480$ denotes evaluation records in the common (q, t_q, r_q) versioned-corpus format; the source temporal expressions and answer targets are inherited from TempQuestions, while record-level issue times and candidate histories instantiate the retrieval setting evaluated here. We reserve the official TimeQA Easy and Hard splits for the clearly off-the-shelf external evaluation. Controlled streaming, conflict, and trend stress tests use evaluation partitions that are time-disjoint from the StreamingQA-derived partitions supplying router labels and remain supporting evidence rather than external-transfer benchmarks (Liska et al., 2022). Across all corpora, indices are frozen at t_q, MinHash removes near-duplicate passages before splitting, and construction-time checks verify supersession identity, effective time, and answer-bearing spans.

A.5 Baseline Implementation Contracts

Naive RAG retrieves 20 Contriever passages and reranks to 10 without temporal features. Version filter keeps the latest observable member of each supersession group before retrieval, whereas Fixed window (7d) applies the same seven-day lookback to every query. TimeAwareRerank adds recency and query-time compatibility to the cross-encoder score after retrieval, while Conflict-aware rerank adds contradiction probability at the same stage. Multi-round temporal RAG follows the iterative verification structure of Madam-RAG (Wang et al., 2025); RouterEns+NLI predicts retrieval need and uses NLI conflict handling without changing pre-ranking eligibility. The capacity stress test uses Llama-3.1-8B-Instruct at 128k, once with evidence matched to time-agnostic RAG and again with compute-matched and best-effort coverage expansion. TCDE-Admission conditions a dual-encoder query representation on requested time and window, trains target-supporting passages against incompatible superseded and future hard negatives, and searches a query-time-admitted partition. All locus-controlled methods preserve the shared Llama-2 generator, reranker, top-10 generator evidence, greedy decoding, and metric implementation; the separately identified 128k rows change generator and context capacity together. Baseline thresholds and temporal weights are selected on the same held-out calibration partition used for operating-point selection and are not retuned on headline or TimeQA test sets. Ret/Q counts dense search invocations, including each multihop call, while the fixed top-5 BM25 probe contributes to measured latency but not to Ret/Q.

A.6 Extended Result Tables

The following tables give the public rerun, full headline and controlled-stress-test results, threshold frontier, request-level outcomes, mechanism decomposition, evaluator audits, locus-matched comparisons, independent external tests, and cross-stack robustness.

Public-evidence sensitivity check. ChronoGate is recomputed on the public-evidence subsets of WikiRevision-Real, GovPolicy-Open, and TemporalQuestions-derived; the other rows are repeated headline references rather than independent reruns. The recomputed ChronoGate cells shift by at most 0.3 points, so this table supports sensitivity of the selected system but is not used as an all-method replication claim.

Full benchmark ledger. The controlled stress suite and each headline corpus show the same method ordering. The full table also makes the distinction between answer correctness, temporal support, and evidence faithfulness visible rather than compressing all three into TC.

Mechanism decomposition. ChronoGate reduces all three disjoint error categories relative to RouterEns+NLI, with the largest absolute reduction in missing valid support. The component rows then identify temporal windowing as the main source of that improvement while preserving a distinct decoder-side faithfulness effect.

Evaluator cross-check. The 800-query validation pool separates scorer agreement from the independent 1,800-query method audit below. SlotFaith improves agreement on numeric questions, where exact values create the largest disagreement for the NLI evaluator.

Table 4: Public-evidence sensitivity check on WikiRevision-Real, GovPolicy-Open, and TemporalQuestions-derived. Only the ChronoGate rows are recomputed on the public-evidence subsets; comparator rows repeat the headline values as fixed references and must not be interpreted as independent reruns. Small ChronoGate shifts reflect subset-specific preprocessing. Long-context uses the separate Llama-3.1 128k capacity stack; the remaining methods use the shared Llama-2 generator.

Method	TC (%)	Fresh (%)	Fth (%)	Ret/Q
WikiRevision-Real (public-source corpus; n=48,612)				
Naive RAG	68.2 ± 0.9	64.7 ± 1.0	71.3 ± 0.8	1.00
Version filter	77.1 ± 0.6	75.2 ± 0.6	79.0 ± 0.6	1.00
Conflict-aware rerank	79.4 ± 0.6	77.2 ± 0.6	80.8 ± 0.5	1.00
Multi-round temporal RAG	79.8 ± 0.6	78.4 ± 0.6	81.6 ± 0.5	2.14
RouterEns+NLI	81.2 ± 0.5	80.4 ± 0.5	82.9 ± 0.5	0.96
Long-context (128k)	80.7 ± 0.5	79.1 ± 0.6	82.4 ± 0.5	1.00
ChronoGate	88.1 ± 0.4	86.8 ± 0.4	89.0 ± 0.4	0.68
GovPolicy-Open (n=22,840)				
Naive RAG	74.8 ± 0.8	72.1 ± 0.9	76.4 ± 0.7	1.00
Version filter	79.4 ± 0.5	77.8 ± 0.6	80.9 ± 0.5	1.00
Conflict-aware rerank	81.3 ± 0.6	79.6 ± 0.6	82.7 ± 0.5	1.00
Multi-round temporal RAG	81.2 ± 0.5	80.1 ± 0.5	82.4 ± 0.5	2.08
RouterEns+NLI	83.6 ± 0.5	82.8 ± 0.5	84.8 ± 0.5	0.93
Long-context (128k)	82.9 ± 0.5	81.4 ± 0.5	84.2 ± 0.5	1.00
ChronoGate	89.8 ± 0.4	89.0 ± 0.4	91.1 ± 0.3	0.74
TemporalQuestions-derived (n=12,480 records)				
Naive RAG	73.4 ± 1.1	70.8 ± 1.2	75.2 ± 1.0	1.00
Version filter	80.2 ± 0.6	78.4 ± 0.7	81.6 ± 0.5	1.00
Conflict-aware rerank	81.7 ± 0.7	79.4 ± 0.8	82.9 ± 0.6	1.00
Multi-round temporal RAG	81.9 ± 0.6	80.4 ± 0.6	83.2 ± 0.5	2.11
RouterEns+NLI	84.1 ± 0.5	83.2 ± 0.5	85.4 ± 0.5	0.94
Long-context (128k)	83.2 ± 0.6	81.8 ± 0.6	84.6 ± 0.5	1.00
ChronoGate	91.0 ± 0.4	89.8 ± 0.4	91.8 ± 0.3	0.69

ChronoGate shifts by only 0.3 points on each public-evidence subset. Because the comparator cells are repeated references, this table establishes sensitivity of the ChronoGate row to evidence provenance; the independent cross-method transfer evidence is provided by the official TimeQA evaluation.

Table 5: Full results across the headline and supporting benchmarks. The controlled stress tests remain supporting evidence, while the four headline datasets carry the main claim in the body of the paper. All intervals are 95% bootstrap confidence intervals over queries. The long-context rows use Llama-3.1-8B-Instruct and are capacity stress tests; other rows use the locus-controlled Llama-2 stack.

Method	TC (%)	Fresh (%)	Fth (%)	Ret/Q
Controlled streaming, conflict, and trend stress tests ($n = 127,420$)				
No-RAG (parametric)	61.3 ± 0.8	58.7 ± 0.9	64.2 ± 0.7	0.00
Naive RAG	72.4 ± 0.6	69.8 ± 0.7	74.6 ± 0.5	1.00
Version filter	78.9 ± 0.5	76.2 ± 0.6	79.4 ± 0.5	1.00
Conflict-aware rerank	82.6 ± 0.4	80.1 ± 0.5	83.2 ± 0.4	1.00
Multi-round temporal RAG	81.4 ± 0.5	79.6 ± 0.5	82.8 ± 0.4	2.18
Long-context (32k)	79.8 ± 0.5	78.4 ± 0.5	81.7 ± 0.4	1.00
Long-context (128k)	84.1 ± 0.4	82.3 ± 0.4	85.4 ± 0.4	1.00
RouterEns+NLI	83.4 ± 0.4	82.8 ± 0.4	84.9 ± 0.4	0.94
ChronoGate	89.7 ± 0.3	88.4 ± 0.3	90.2 ± 0.3	0.71
WikiRevision-Real (10,284 entities, avg. 14.8 edits/year; n=48,612)				
No-RAG (parametric)	54.8 ± 1.2	51.3 ± 1.3	58.4 ± 1.1	0.00
Naive RAG	68.2 ± 0.9	64.7 ± 1.0	71.3 ± 0.8	1.00
Version filter	77.1 ± 0.6	75.2 ± 0.6	79.0 ± 0.6	1.00
Fixed window (7d)	74.6 ± 0.7	71.8 ± 0.8	76.2 ± 0.7	1.00
Conflict-aware rerank	79.4 ± 0.6	77.2 ± 0.6	80.8 ± 0.5	1.00
Multi-round temporal RAG	79.8 ± 0.6	78.4 ± 0.6	81.6 ± 0.5	2.14
Long-context (128k)	80.7 ± 0.5	79.1 ± 0.6	82.4 ± 0.5	1.00
RouterEns+NLI	81.2 ± 0.5	80.4 ± 0.5	82.9 ± 0.5	0.96
ChronoGate	88.4 ± 0.4	87.1 ± 0.4	89.3 ± 0.4	0.68
GovPolicy-Open (5,127 policy documents; n=22,840)				
Naive RAG	74.8 ± 0.8	72.1 ± 0.9	76.4 ± 0.7	1.00
Version filter	79.4 ± 0.5	77.8 ± 0.6	80.9 ± 0.5	1.00
Conflict-aware rerank	81.3 ± 0.6	79.6 ± 0.6	82.7 ± 0.5	1.00
Multi-round temporal RAG	81.2 ± 0.5	80.1 ± 0.5	82.4 ± 0.5	2.08
Long-context (128k)	82.9 ± 0.5	81.4 ± 0.5	84.2 ± 0.5	1.00
RouterEns+NLI	83.6 ± 0.5	82.8 ± 0.5	84.8 ± 0.5	0.93
ChronoGate	90.1 ± 0.4	89.2 ± 0.4	91.4 ± 0.3	0.74
FinNews-Exec (8,412 executive changes, 2020–2024; n=31,548)				
Naive RAG	71.6 ± 0.9	68.4 ± 1.0	73.8 ± 0.8	1.00
Version filter	78.3 ± 0.6	76.4 ± 0.7	79.8 ± 0.6	1.00
Conflict-aware rerank	80.2 ± 0.6	78.1 ± 0.7	81.4 ± 0.6	1.00
Multi-round temporal RAG	80.4 ± 0.6	79.2 ± 0.6	81.9 ± 0.5	2.21
Long-context (128k)	81.8 ± 0.5	80.2 ± 0.6	83.1 ± 0.5	1.00
RouterEns+NLI	82.4 ± 0.5	81.6 ± 0.5	83.7 ± 0.5	0.95
ChronoGate	89.2 ± 0.4	88.3 ± 0.4	90.6 ± 0.3	0.72
TemporalQuestions-derived (author-constructed records; n=12,480)				
Naive RAG	73.4 ± 1.1	70.8 ± 1.2	75.2 ± 1.0	1.00
Version filter	80.2 ± 0.6	78.4 ± 0.7	81.6 ± 0.5	1.00
Conflict-aware rerank	81.7 ± 0.7	79.4 ± 0.8	82.9 ± 0.6	1.00
Multi-round temporal RAG	81.9 ± 0.6	80.4 ± 0.6	83.2 ± 0.5	2.11
Long-context (128k)	83.2 ± 0.6	81.8 ± 0.6	84.6 ± 0.5	1.00
RouterEns+NLI	84.1 ± 0.5	83.2 ± 0.5	85.4 ± 0.5	0.94
ChronoGate	91.3 ± 0.4	90.1 ± 0.4	92.1 ± 0.3	0.69

Across the controlled suite and all headline corpora, ChronoGate improves TC, freshness, and faithfulness simultaneously. The fixed-window and fixed-version rows show the cost of query-independent admission; the RouterEns+NLI rows isolate the strongest adaptive post-admission alternative.

Table 6: Threshold-to-budget mapping for the reference stack, together with transfer across domains. End-to-end latency was measured on a single A100-based implementation; these stack-specific values should be compared within this paper and should not be interpreted as hardware-independent constants. The transfer rows reuse the calibrated operating point $\theta = 0.5$, with no retuning for individual domains.

Setting	Latency (ms)	Ret/Q	TC (%)	Fresh (%)
Threshold-to-budget mapping (WikiRevision-Real, calibrated router)				
$\theta = 0.3$ (conservative)	261	0.92	90.1 ± 0.4	89.3 ± 0.4
$\theta = 0.5$ (balanced)	238	0.68	88.4 ± 0.4	87.1 ± 0.4
$\theta = 0.7$ (aggressive)	198	0.41	84.2 ± 0.6	82.8 ± 0.6
$\theta = 0.9$ (minimal retrieval)	156	0.18	76.8 ± 0.9	74.6 ± 1.0
Cross-domain threshold transfer (calibrated router, $\theta = 0.5$)				
GovPolicy-Open	247	0.74	90.1 ± 0.4	89.2 ± 0.4
FinNews-Exec	244	0.72	89.2 ± 0.4	88.3 ± 0.4
TemporalQuestions-derived	239	0.69	91.3 ± 0.4	90.1 ± 0.4
Cross-domain transfer without calibration (same $\theta = 0.5$)				
GovPolicy-Open	312	0.91	86.4 ± 0.6	84.8 ± 0.7
FinNews-Exec	178	0.32	79.1 ± 0.8	77.4 ± 0.9
TemporalQuestions-derived	284	0.82	85.7 ± 0.6	84.2 ± 0.7

Table 7: Compact mechanism summary on WikiRevision-Real. Panel A reports the disjoint support-error decomposition. Panel B reports the matched-budget component checks that separate temporal windowing from calibration and the conflict gate. Temporal windowing accounts for the largest component loss; calibration and conflict gating contribute smaller, distinct effects.

Panel A: Error decomposition

Method	E_C (%)	E_P (%)	E_R (%)	Total (%)
RouterEns+NLI	6.2	5.9	6.7	18.8
Long-ctx 128k	5.1	8.7	5.5	19.3
ChronoGate	3.2	4.1	4.3	11.6

Panel B: Component summary

Setting	TC	Fresh	Fth	Ret/Q
Full system	88.4	87.1	89.3	0.68
− Temp. window	79.8	78.4	81.2	0.68
− Calibration	84.8	83.5	85.9	0.68
− Conflict gate	86.1	84.8	84.2	0.68

Table 8: Evaluator audit for the 800-query human-validation pool. Human majority labels and SlotFaith preserve the headline method ordering, with numeric questions producing the largest evaluator disagreement.

Check	Overall (%)	Numeric (%)	High-conflict (%)
NLI-based TC vs. human TC	91.2	82.7	87.4
SlotFaith vs. human labels	90.8	86.2	—

Table 9: Source-separated FinNews validation on 12,000 queries grounded only in SEC filings and company releases. The post-retrieval temporal reranker and ChronoGate use the shared reference stack; intervals are query-record bootstrap 95% confidence intervals.

Method	TC	Fresh	Fth	Ret/Q	TC-positive
TimeAwareRerank	80.6 ± 0.6	79.0 ± 0.6	82.0 ± 0.5	1.00	—
ChronoGate	88.9 ± 0.4	88.0 ± 0.4	90.2 ± 0.3	0.72	10,668

Source separation retains the pre-ranking advantage. On the open FinNews crosswalk, ChronoGate adds 8.3 TC points over post-retrieval TimeAwareRerank while using fewer dense searches.

Table 10: Targeted hard-slice subsets within the 800-query human audit. Panel A reports majority-vote human TC/Fth for the listed subsets; the full pool contributes to the aggregate evaluator audit in Table 8. Panel B reports non-NLI agreement with human labels and paired ChronoGate TC outcomes.

Slice	n	ChronoGate	RouterEns+NLI	Long-context 128k
Numeric	186	85.5 / 86.8	79.3 / 80.4	78.1 / 79.0
Date	141	87.2 / 88.1	81.0 / 81.8	80.4 / 81.1
High-conflict	152	83.7 / 84.5	76.5 / 78.1	72.9 / 74.0

Slice	Non-NLI agreement	vs. Router wins / ties / losses	vs. Long-context wins / ties / losses
Numeric	86.2%	92 / 58 / 36	101 / 47 / 38
Date	87.1%	74 / 37 / 30	78 / 31 / 32
High-conflict	88.6%	93 / 27 / 32	102 / 19 / 31

Hard-slice judgments track temporal consistency directly. The human-majority comparisons favor ChronoGate on numeric, date, and high-conflict questions, and the non-NLI check agrees with human labels on 86.2–88.6% of these subsets. The paired counts show that the aggregate ordering is not produced by a small collection of extreme examples.

Table 11: Coverage-expanded long-context counterfactual on WikiRevision-Real. Compute-matched and best-effort settings expand retrieved coverage before Llama-3.1-8B-Instruct 128k generation; the final two columns expose conflict-sensitive outcomes and deployment cost. ChronoGate uses the Llama-2 reference stack, so this table is a capacity stress test rather than a matched-generator locus comparison.

Setting	TC	Fresh	Fth
Matched-evidence 128k	80.7 ± 0.5	79.1 ± 0.6	82.4 ± 0.5
Coverage-optimized, compute-matched	83.6 ± 0.5	81.9 ± 0.6	84.7 ± 0.5
Coverage-optimized, best-effort	86.1 ± 0.5	84.8 ± 0.5	86.9 ± 0.5
ChronoGate	88.4 ± 0.4	87.1 ± 0.4	89.3 ± 0.4

Setting	Tokens	P50/P99	High-conflict TC	E_P
Matched-evidence 128k	22,684	691/902	72.1 ± 0.9	8.7
Coverage-optimized, compute-matched	24,190	712 / 925	76.8 ± 0.8	7.1 ± 0.6
Coverage-optimized, best-effort	35,420	934 / 1,258	79.4 ± 0.8	6.2 ± 0.5
ChronoGate	2,947	238/301	84.6 ± 0.7	4.1

Capacity and admission improve different support properties. Coverage expansion steadily raises long-context TC, yet the best-effort setting retains a 6.2% purity error and 79.4 TC under high conflict. ChronoGate reaches 4.1% purity error and 84.6 high-conflict TC with substantially fewer effective tokens.

TCDE makes temporal retrieval competitive without erasing the action-selection effect. Its per-corpus deficit is 1.1–1.2 TC points, and its 27,081 paired wins confirm that it is a demanding

Table 12: Cross-generator and cross-retriever robustness on WikiRevision-Real. Each pair changes one stack component while preserving the admission protocol and evaluation.

Changed component	Setting	Router TC	ChronoGate TC	ChronoGate Fresh	ChronoGate Fth
Generator	Mistral-7B-Instruct	82.8 ± 0.5	89.6 ± 0.4	88.4 ± 0.4	90.8 ± 0.3
Generator	Llama-3-8B-Instruct	84.1 ± 0.4	90.8 ± 0.3	89.7 ± 0.4	91.9 ± 0.3
Retriever	BM25	78.4 ± 0.6	85.8 ± 0.5	84.4 ± 0.5	87.2 ± 0.4
Retriever	ColBERTv2	83.6 ± 0.5	90.2 ± 0.3	89.1 ± 0.4	91.4 ± 0.3

Table 13: Retriever-specific same-locus comparison. TCDE-Admission trains a time/window-conditioned dual encoder with superseded hard negatives and searches only the query-time-valid partition; all downstream components match ChronoGate. Intervals are 95% query-bootstrap intervals.

Dataset	n	TC	Fresh	Fth
WikiRevision-Real	48,612	87.2 ± 0.5	86.2 ± 0.5	88.3 ± 0.5
GovPolicy-Open	22,840	89.0 ± 0.4	88.1 ± 0.5	89.9 ± 0.4
FinNews-Exec	31,548	88.1 ± 0.5	87.2 ± 0.5	89.4 ± 0.4
TemporalQuestions-derived	12,480	90.2 ± 0.4	89.4 ± 0.4	91.1 ± 0.4

Dataset	Ret/Q	P50	TC-positive
WikiRevision-Real	0.80	254	42,390
GovPolicy-Open	0.80	251	20,328
FinNews-Exec	0.82	256	27,794
TemporalQuestions-derived	0.79	247	11,257

Very-high-update TC ($n = 12,153$)	TCDE 79.0 ± 0.9; ChronoGate 81.4 ± 0.9
High-conflict TC ($n = 8,687$)	TCDE 82.2 ± 0.8; ChronoGate 84.6 ± 0.7
Wiki purity error	TCDE 4.5 ± 0.4; ChronoGate 4.1
Paired TC, ChronoGate vs. TCDE	28,399 / 60,000 / 27,081 wins / ties / losses

The per-corpus margin is modest but consistent, and it doubles on the two severe slices. The paired excess of 1,318 favorable outcomes and lower Wiki purity error are consistent with a contribution from explicit set and window selection.

comparator rather than a weak post-hoc control. ChronoGate's larger severe-slice margins align with the cases in which a query-specific boundary must distinguish several closely related versions.

Table 14: External TimeQA evaluation with no TimeQA training or calibration. Easy and Hard are the official test splits. ChronoGate, TCDE-Admission, and RouterEns+NLI share the Llama-2 downstream stack; optimized long context uses Llama-3.1-8B-Instruct as a separate capacity stress test. All rows share the reranker, decoding policy, metrics, and query-bootstrap uncertainty protocol.

Split	Method	TC	Fresh	Fth	Ret/Q	TC-positive
Easy ($n = 2,997$)	RouterEns+NLI	82.1 ± 0.8	80.9 ± 0.8	83.5 ± 0.7	0.96	—
	Optimized long-context 128k	83.0 ± 0.7	81.7 ± 0.8	84.2 ± 0.7	1.00	—
	TCDE-Admission	85.6 ± 0.6	84.4 ± 0.7	86.7 ± 0.6	0.81	2,565
	ChronoGate	87.3 ± 0.6	86.2 ± 0.6	88.5 ± 0.5	0.72	2,616
Hard ($n = 3,078$)	RouterEns+NLI	74.9 ± 0.9	73.2 ± 1.0	76.4 ± 0.9	0.97	—
	Optimized long-context 128k	76.1 ± 0.9	74.6 ± 0.9	77.3 ± 0.8	1.00	—
	TCDE-Admission	78.5 ± 0.8	77.0 ± 0.8	79.8 ± 0.7	0.82	2,416
	ChronoGate	81.2 ± 0.7	79.7 ± 0.8	82.4 ± 0.7	0.73	2,499

ChronoGate leads on both official splits without TimeQA training or calibration. Its margin over TCDE increases from 1.7 points on Easy to 2.7 points on Hard while retaining fewer dense searches.

TimeQA tests transfer beyond the constructed headline corpora. Without TimeQA training or calibration, ChronoGate adds 1.7 points over TCDE on Easy and 2.7 points on Hard. The Hard split also shows the largest margin over optimized long context, consistent with its implicit temporal reasoning requirement.

Independent labels preserve both answer and evidence ordering. Across all three 600-query slices, human temporal consistency and faithfulness rank ChronoGate first. The non-NLI evaluator has higher human agreement than the default NLI evaluator on numeric and date queries and remains directionally consistent under high conflict.

B Additional Reliability and Request-Level Analysis

Timing metadata is a measured input, not an oracle. The main evaluation freezes each index at issue time and derives validity from publication, effective-time, and supersession records. Table 16 perturbs that information directly: Gaussian noise shifts every document time, random masking creates an explicit unknown-time state, and adaptive injection assigns plausible recent times to stale passages. These interventions distinguish ordinary metadata degradation from the more targeted error that changes candidate membership.

Degradation follows the support-boundary mechanism. At 30 days of timestamp noise, ChronoGate retains 78.4 TC, compared with 75.9 for TCDE-Admission, 66.2 for RouterEns+NLI, and 48.7 for a fixed window. At 50% missingness, ChronoGate, TCDE-Admission, and RouterEns+NLI reach 75.6, 71.8, and 69.7 TC. The adaptive stale-injection setting is harder because a false recent timestamp moves incompatible evidence inside the eligible set; ChronoGate nevertheless retains 72.6 TC versus 68.4 and 60.8 for the two learned comparators.

Request-level accounting. Table 17 reports the observable quantities behind threshold choice: temporally inconsistent answers, incorrect skips, dense-retrieval calls, prompt tokens, and P50/P99 latency. Moving from $\theta = 0.3$ to 0.9 reduces Ret/Q from 0.92 to 0.18 and P50 latency from 261 to 156 ms, while TC changes from 90.1 to 76.8. The transferred $\theta = 0.5$ operating point yields 88.4 TC at 0.68 Ret/Q; applications can select another displayed point directly from their omission and latency requirements without relying on a paper-specific scalar utility.

Table 15: Disjoint 1,800-query independent audit. Three annotators, blinded to method identity, label temporal correctness and faithfulness from cited evidence spans; entries in Panel A are majority-vote human TC/Fth with 95% intervals. Panel B compares automatic evaluators with human labels and gives paired ChronoGate-versus-TCDE TC counts. The long-context row uses the Llama-3.1 128k capacity stack; the other methods use the locus-controlled Llama-2 stack.

Slice	Method	n	Human TC	Human Fth
Numeric	ChronoGate	600	85.2 ± 2.8	86.5 ± 2.7
	TCDE-Admission		82.4 ± 3.0	83.6 ± 3.0
	RouterEns+NLI		79.1 ± 3.2	80.3 ± 3.2
	Long-context 128k		78.0 ± 3.3	79.1 ± 3.2
Date	ChronoGate	600	87.0 ± 2.7	88.2 ± 2.6
	TCDE-Admission		84.3 ± 2.9	85.4 ± 2.8
	RouterEns+NLI		81.0 ± 3.1	81.9 ± 3.1
	Long-context 128k		80.5 ± 3.2	81.2 ± 3.1
High-conflict	ChronoGate	600	83.5 ± 3.0	84.4 ± 2.9
	TCDE-Admission		81.1 ± 3.1	82.0 ± 3.1
	RouterEns+NLI		76.6 ± 3.4	78.0 ± 3.3
	Long-context 128k		73.1 ± 3.6	74.2 ± 3.5

Slice	NLI–human agreement	Non-NLI–human agreement (Wilson 95% CI)
Numeric	83.1%	89.2% [86.5%, 91.4%]
Date	85.8%	91.0% [88.5%, 93.0%]
High-conflict	87.0%	88.7% [85.9%, 91.0%]

Slice	ChronoGate vs. TCDE TC wins / ties / losses
Numeric	153 / 311 / 136
Date	142 / 332 / 126
High-conflict	174 / 266 / 160

Human temporal consistency and faithfulness place ChronoGate first on all three slices. The paired counts favor ChronoGate in each slice, and the separate non-NLI scorer preserves the ordering with 88.7–91.0% agreement to human labels.

Table 16: Timestamp-corruption and adaptive-injection stress tests on WikiRevision-Real. Noise perturbs 48,612 queries / 3,184,760 document timestamps, missingness perturbs 48,612 queries / 3,184,760 document timestamps, and adaptive injection evaluates 48,612 queries / 243,060 inserted stale-document placements. Entries are TC with query-bootstrap 95% intervals.

Timestamp noise	0 days	7 days	14 days	30 days
ChronoGate	88.4 ± 0.4	86.9 ± 0.5	83.7 ± 0.6	78.4 ± 0.8
TCDE-Admission	87.2 ± 0.5	85.1 ± 0.6	81.6 ± 0.7	75.9 ± 0.9
RouterEns+NLI	81.2 ± 0.5	78.5 ± 0.6	73.7 ± 0.8	66.2 ± 1.0
Fixed window	74.6 ± 0.7	69.8 ± 0.8	60.9 ± 1.0	48.7 ± 1.2

Missing timestamps	0%	10%	25%	50%
ChronoGate	88.4 ± 0.4	86.8 ± 0.5	82.9 ± 0.7	75.6 ± 0.9
TCDE-Admission	87.2 ± 0.5	84.9 ± 0.6	80.1 ± 0.8	71.8 ± 1.0
RouterEns+NLI	81.2 ± 0.5	79.8 ± 0.6	76.4 ± 0.8	69.7 ± 1.0

Method under 50% adaptive stale injection	TC
ChronoGate	72.6 ± 0.9
TCDE-Admission	68.4 ± 1.0
RouterEns+NLI	60.8 ± 1.1
Naive RAG	36.2 ± 1.2

Table 17: Request-level quality–cost frontier on WikiRevision-Real ($n = 48,612$). Stale and incorrect-skip columns are raw error counts; tokens are effective prompt tokens. All latency values use the single-A100 reference stack.

Setting	TC	Stale answers	Incorrect skips	Ret/Q	Tokens	P50/P99
ChronoGate $\theta = 0.3$	90.1	2,187	923	0.92	3,520	261/329
ChronoGate $\theta = 0.5$	88.4	2,418	2,042	0.68	2,947	238/301
ChronoGate $\theta = 0.7$	84.2	2,789	3,650	0.41	2,260	198/259
ChronoGate $\theta = 0.9$	76.8	3,016	6,834	0.18	1,640	156/211
TCDE-Admission	87.2 ± 0.5	2,936	2,108	0.80	3,280	254/322
RouterEns+NLI	81.2	5,350	1,100	0.96	4,740	276/342
Optimized long-context 128k	86.1 ± 0.5	4,100	0	1.00	35,420	934 / 1,258