HISTORICAL CASE CLOSURE FOR CLINICAL GUIDANCE: SCORING COMPLETE CUTOFF-TIME EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A reviewer reopening a dosing decision made before a later label revision can recover the right recommendation and still fail to close the case: the answer may point to the successor revision, omit co-active guidance, or cite a span that cannot be resolved in the claimed source. We model this handoff as an evidence-closure path from a dated request to its scoped claim, cutoff-valid source events, reconstructed state, and revision-bound witnesses. Historical case closure (HCC) scores the whole path, accepts adjudicated alternatives, requires coverage for absence claims, and records the first broken edge. On a common 359-request clinical-guidance manifest, GuidanceTrace closes 359/359 requests (95% Wilson CI: 98.9–100%), a bitemporal document retrieve–rerank–span system averages 90.3% $\pm$ 0.7 across five seeds, and RAG+TimeFilter closes 42.1%. A matched keying intervention supplies the mechanism test: on separate 2,000-event and 1,200-case monitoring tasks it changes drift and contradiction F1 by at most 0.3 points, yet on the closure manifest it loses 52 cases (85.5% (307/359); paired deficit 14.5 points [10.9, 18.4]), including 37 first failures at effective-time version or bundle selection. Independent second-site replay preserves the advantage—92.2% (166/180) closure versus 81.7% (147/180)—and lowers median review time from 31.6 to 6.8 minutes. HCC exposes source-time addressability and evidence binding as concrete design targets for reviewer-complete historical answers.

1 INTRODUCTION

Reopen a dosing decision at cutoff τ, before a later label revision. The reviewer must follow one evidence path: recover the claim for the relevant population, enumerate every guideline or label event active at τ, reconstruct the state seen then, and inspect the supporting text in those exact revisions. A present-day answer can match the recommendation's semantics yet branch to a successor source, drop a co-active event, or terminate at an unverifiable span. In each case the answer sounds plausible while the historical review remains open. Living evidence workflows Elliott et al. (2017), post-marketing withdrawals and safety surveillance Onakpoya et al. (2016); Bate & Evans (2009), and national surveillance streams Platt et al. (2018) make this path a recurring clinical-governance object.

This path yields a direct mechanism test. GuidanceTrace keeps two addresses: `claim_id` aligns a recommendation across wording and sources, while `assertion_id` retains each source-time event. We replace the latter lookup key with the former while holding extraction, scope, provenance, snapshots, thresholds, manifest, and scorer fixed. On the separate 2,000-event drift and 1,200-case contradiction tasks, F1 changes by at most 0.3 points. On the 359-request historical manifest, however, closure falls from 359/359 to 85.5% (307/359); effective-time version or bundle selection accounts for 37 of the 52 newly open cases. The intervention pinpoints source-event addressability as information that semantic matching does not retain for this selector.

Historical case closure (HCC) turns the review path into the scored unit. Each request fixes its eligible sources and cutoff; each gold object is a set of complete paths that can accommodate co-active recommendations and equivalent evidence. A submission closes only when its answer, effective events, serialized state, and revision-bound witnesses belong to one admissible path. Absence claims must cover the eligible revision or source set, and every non-closing output is assigned to its first broken edge. This output interface lets assertion stores, temporal databases, document retrievers, and generated answers meet the same reviewer-facing criterion.

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.

The neighboring literatures score nodes of this path. Temporal and streaming QA evaluate answer selection as knowledge changes Zhang & Choi (2021); Liška et al. (2022); Shang et al. (2022); temporal RAG adds outdated-document and multi-document reasoning tests Ouyang et al. (2025); Chen et al. (2025); citation metrics score support Gao et al. (2023); Rashkin et al. (2023); and temporal databases reconstruct valid-time state Dignös et al. (2012); Kaufmann et al. (2013). Biomedical knowledge graphs and living-guidance workflows organize the underlying semantic and revision structure Chandak et al. (2023); Breit et al. (2020); Elliott et al. (2017). HCC scores whether these nodes connect into a complete dated-review path, including multi-event and negative cases. Appendix Table S4 gives the predicate-level comparison.

The common manifest spans 142 negative and 73 hard requests. GuidanceTrace closes all 359, TemporalSpanQA averages 90.3% $\pm$ 0.7 closure over five seeds, and RAG+TimeFilter closes 42.1%. Equal totals need not imply equal systems: two controls at 76.3% fail predominantly at evidence binding versus version/state reconstruction. Source-chain-isolated evaluation and blinded span review test the path beyond the primary source families. At an independent second site, GuidanceTrace closes 92.2% (166/180) requests versus 81.7% (147/180) for TemporalSpanQA, with median review time of 6.8 versus 31.6 minutes.

The contribution is a reviewer-complete historical output object whose score follows the evidence path rather than stopping at answer correctness. The manifest makes that object comparable across system families; the matched key intervention identifies why one path breaks; and the failure ledger converts each open case into a concrete modeling target. GuidanceTrace provides a reference realization with 94.2% drift F1 and 94.0% contradiction F1 on the separate monitoring tasks, while the closure and second-site results establish the consequence for historical review.

2 Methods and Experimental Setup

2.1 Historical Case Closure as a Scored Reviewer Handoff

A case $c = (q, \tau, \mathcal{U}_c, \mathcal{G}_c)$ fixes a request q, cutoff τ, eligible source-event universe $\mathcal{U}_c$, and a nonempty set of adjudicated gold bundles $\mathcal{G}_c$. Each $g = (H, E, W, Z, R) \in \mathcal{G}_c$ contains a set of ontology-normalized scoped claims H, the source events E needed to justify them, evidence witnesses W, a canonical point-in-time state Z, and a coverage record R. Sets, rather than one canonical claim or event, preserve co-active recommendations and adjudicated semantic equivalents without accepting arbitrary alternatives.

A source event is $e = (\texttt{source_id}, \texttt{source_version}, t_{\text{pub}}, t_{\text{from}}, t_{\text{to}}, \texttt{status})$. It is active when $t_{\text{from}} \leq \tau < t_{\text{to}}$. An explicit effective date determines t_{from}; otherwise publication time is used. A superseding revision closes its predecessor's half-open interval. Corrections insert a new event and close the corrected event at the correction's effective time; withdrawals insert a record with `status=withdrawn` whose state acts as a tombstone. Transaction time is retained for provenance but never substituted for effective time. Simultaneous non-equivalent events remain jointly active; equivalent alternatives are separate members of $\mathcal{G}_c$, so the scorer applies no arbitrary timestamp tie break.

Every witness identifies an `assertion_id` and `revision_digest`, then carries `offset_start`, `offset_end`, and `span_digest`. The half-open character interval $[\texttt{offset_start}, \texttt{offset_end})$ is applied to the scorer's normalized revision text, and `span_digest` commits to the resulting bytes. These fields bind one exact span to one immutable event. Z is the UTF-8 canonical JSON serialization of the sorted active records. Each record carries the claim and assertion identifiers, source and source version, effective interval, status, and revision digest. Arrays sort lexicographically by `claim_id`, `source_id`, effective time, and `source_version`; timestamps use UTC RFC 3339; absent fields are explicit `null`; and duplicate records are invalid. The top-level object has exactly five typed members—claim set H, source-event set E, witness set W, serialized state Z, and coverage record R—so a system returns the same portable object $\hat{y} = (\hat{H}, \hat{E}, \hat{W}, \hat{Z}, \hat{R})$, independently of its internal architecture.

For one jointly admissible g, the four predicates are

$$A_{\text{sem}} = \mathbb{1}[\text{norm}(\hat{H}) = H], \qquad A_{\text{src}} = \mathbb{1}[\text{eligible}(\hat{E}, \hat{R}, \tau) = (E, R)],$$
$$A_{\text{state}} = \mathbb{1}[\text{canon}(\hat{Z}) = Z], \qquad A_{\text{span}} = \mathbb{1}[\text{verify}(\hat{W}, \hat{E}, W) = 1]. \tag{1}$$

The witness verifier requires every returned witness to resolve to its claimed event and digest, every claim to have its required witnesses, and the submitted witness set to equal an admissible set after removal of byte-identical duplicates. Case closure and the aggregate metric are

$$C(c, \hat{y}) = \mathbb{1}[\exists g \in \mathcal{G}_c : A_{\text{sem}} A_{\text{src}} A_{\text{state}} A_{\text{span}} = 1], \qquad \text{HCC} = |M|^{-1} \sum_{c \in M} C(c, \hat{y}_c).$$

The existential is over complete gold bundles; components from different alternatives cannot be mixed.

As a schematic scoring example, let one admissible positive bundle require co-active events $\{e_1, e_2\}$ and witnesses $\{w_1, w_2\}$. Returning the correct H and Z with only e_1 fails A_{src} at stage 1; returning both events with w_2 bound to a successor revision passes source selection but fails A_{span} at stage 3. For a negative case with eligible revisions $\{r_1, r_2, r_3\}$, `NO_CHANGE` closes only when R commits to the active-claim digest at all three revisions. The same predicate order therefore evaluates positive, multi-event, and absence claims without method-specific scoring.

For positive cases, R has `coverage_type=POSITIVE` and an empty entry array because required-event completeness is carried by E and W; positive bundles may require several events or spans. For `NO_CHANGE`, R must enumerate every eligible revision in the requested interval and commit to the canonical active-claim digest at each transition. For `NO_CONFLICT`, it must enumerate every eligible source in the prespecified source set and commit to each active scoped claim at τ. Equality to the frozen coverage set makes a single convenient span insufficient evidence of absence. Empty answers and unsupported abstentions fail.

Two clinical informaticists construct $\mathcal{G}_c$ directly from frozen source revisions, adjudicating scoped claims, co-active events, equivalent passages, effective intervals, and negative-case coverage; a pharmacist resolves disagreements. The rubric, source eligibility, gold bundles, and manifest are frozen before any comparator output is inspected. The scorer maps every non-closing output to one exhaustive first-failure stage: (1) historical-version or coverage error; (2) answer/bundle error, including malformed, missing, non-terminating, semantically mismatched, or mutually incompatible output; (3) unverifiable, incomplete, or non-unique evidence; or (4) state reconstruction. Syntactically invalid output enters stage 2 before field-level checks; otherwise stages use the displayed order. This taxonomy preserves the all-or-nothing endpoint while identifying the earliest reviewer-blocking defect.

2.2 GuidanceTrace Output Representation

GuidanceTrace instantiates the contract with immutable, scope-aware assertions. Each record contains subject, predicate, object, condition, disease stage, line of therapy, age band, comorbidity flags, source, publication and effective times, status, confidence, revision digest, and span digest. LayoutLMv3 document parsing, PubMedBERT/BioBERT entity extraction, BioRED relation extraction, and MetaMap ontology grounding produce the record Huang et al. (2022); Gu et al. (2022); Lee et al. (2020); Luo et al. (2022); Aronson & Lang (2010); Wang et al. (2018); UMLS, SNOMED CT, and RxNorm provide concept normalization Bodenreider (2004); Rosenbloom et al. (2009); Liu et al. (2005). Component and source-level extraction scores appear in Appendix Tables S12–S13.

Two identifiers enforce complementary invariants. `claim_id` is the SHA-256 digest of subject, predicate, object, and scope, aligning semantically equivalent recommendations across revisions. `assertion_id` hashes `claim_id`, `source_id`, `source_version`, and effective interval, retaining each source-time event as a separately addressable object. Across 78.4M assertions, claim matching aligns 11.3% across sources; 189 of 200 manually reviewed matches are true semantic duplicates (94.5%). Revision and span digests bind the event to the source bytes returned to the scorer. Append-only logs and snapshot pointers reconstruct the state active at a cutoff without full-corpus replay, following provenance and temporal-data design principles Glavic (2021); Psallidas & Wu (2018); Overeem et al. (2021); Armbrust et al. (2020); Kaufmann et al. (2013); Dignös et al. (2012). The resulting output preserves semantic alignment, source-event addressability, effective-time reconstruction, and independently recomputable evidence in one scorer-facing unit.

2.3 Monitoring Tasks and Trust Boundary

Drift detection compares aligned assertions across successive versions and assigns *strengthened*, *weakened*, *reversed*, or *expanded subgroup*. Contradiction detection applies 61 frozen templates to active assertions under matched scope. The same scope tuple serves both tasks, so experiments test a shared evidence representation instead of task-specific entity matching. For the separate-task F1 columns reported for document systems, TemporalSpanQA and Baseline F receive the same successive-version or cross-source document pairs and cutoff metadata, then emit the frozen drift class or contradiction decision through their constrained decoder; task F1 scores that label before the HCC adapter constructs a closure bundle. International transfer freezes the representation, rule inventory, and all thresholds before ESC/EMA and JCS/PMDA cases are inspected.

The system is tamper-evident under a trusted-operator, authenticated-write threat model and detects content, effective-time, and snapshot inconsistencies within that boundary. The reported SHA-256 digests are content commitments; authenticity comes from the authenticated append log and frozen revision registry. The evaluated attacks modify registered content or temporal state while holding that trust root fixed. The main observed failure modes are complex subgroup scope, cross-sentence relations, transferred terminology, and rule maintenance burden.

2.4 Data, Baselines, and Protocols

We curated 3,842 guideline versions from 27 specialty societies, 12,417 DailyMed label updates, and 4.8M de-identified medication and laboratory events. The main tasks contain 2,000 guideline-change events from 11 families and 1,200 cross-source contradiction cases. Two clinical informaticists annotated both tasks ($\kappa = 0.87$ drift; 0.84 contradiction), with pharmacist adjudication for contradiction disagreements. Forty percent of drift cases and 45% of contradiction cases are negative. Family and complete source-version chains define the splits: ACC/AHA and NCCN are excluded when used for within-portfolio holdout, and ESC/JCS remain isolated for external drift evaluation.

The 359-request closure manifest is an evaluation-only set sampled and frozen before comparator scoring. Stratification covers request type, source pair, cutoff age, positive/negative status, and easy/moderate/hard difficulty; Appendix Table S5 gives the exact mutually exclusive counts, including 142 negative and 73 hard cases. Gold bundles are constructed from source revisions rather than from GuidanceTrace outputs, and no manifest request or label is used to fit extraction models, rules, thresholds, retrieval depth, or span thresholds. The 2,000-event drift and 1,200-case contradiction benchmarks are separate task evaluations; their F1 values are not manifest-level projections.

Baseline A is a timestamp-sorted property graph. Baseline D removes the persistent revision-to-span digest from the full versioned pipeline; its adapter still emits the required witness by hashing the selected source bytes at submission, so failures reflect incorrect or non-recoverable span selection rather than a missing schema field. Baseline G removes versioning, digests, and snapshots. Baseline E is a PostgreSQL 16.2 bitemporal store over the extracted records. Baseline F uses GPT-4-turbo, Chroma, `text-embedding-3-large`, cutoff filtering, temperature 0, and top-k selected from $\{5, 10, 20, 40\}$. The collapsed-ID intervention holds extracted assertions, scope tuples, provenance fields, digests, append-only log, snapshots, indexes, thresholds, hardware, manifest, and scorer fixed. It replaces the selector's `assertion_id` lookup key with `claim_id`; when several source-time records share that key, the unchanged effective-time ordering exposes one selected record rather than separately addressable co-active candidates. Source records remain in the append log, isolating the tested lookup policy from extraction and storage.

TemporalSpanQA is a bitemporal document retrieve–rerank–span comparator designed to test HCC outside an assertion-store architecture. It indexes the same eligible documents as 384-token chunks with 64-token overlap, applies the same cutoff timestamps, retrieves with `text-embedding-3-large`, reranks with PubMedBERT, selects character-offset spans, and emits the scorer schema through a constrained GPT-4-turbo JSON decoder. Retrieval depth and span threshold are tuned on development families, then frozen. We run five index/model seeds (1701, 2718, 3141, 5772, 8119); the primary summary is mean closure and standard deviation across all five, while seed 1701 (the first listed, not the best-scoring seed) supplies raw request counts and failure accounting. All systems receive identical source eligibility and cutoff information.

The blinded semantic audit samples 400 general GuidanceTrace assertions before inspection; it is independent of the curated closure manifest and therefore assesses a different denominator. Of these, 396 are evaluable and 371 are span-supported, yielding 93.7%; the evaluable strata are 170 guideline, 134 label, and 92 operational assertions. Cohen's $\kappa = 0.84$ is computed on all 400 pre-consensus three-way labels (supported, unsupported, unscorable); four unscorable spans are excluded only from the support-rate denominator. Thus 359/359 manifest closure measures submitted bundles on adjudicated requests, whereas 86.8% extraction F1 and 93.7% span support characterize upstream extraction and general-corpus assertions.

International contradiction transfer contains 320 independently annotated cases: 184 ESC/EMA and 136 JCS/PMDA, with 180 positives and 140 negatives. Complete source-version chains are isolated from development. Two blinded clinical informaticists label the cases, a pharmacist adjudicates disagreements, and the 61 rules and all thresholds remain frozen. The second-site replay similarly freezes 180 historical requests from non-overlapping source chains before scoring GuidanceTrace, TemporalSpanQA, and the site's existing review workflow under the same rubric.

System experiments use a 3-node AMD EPYC 7763 cluster (32 vCPUs, 256 GiB RAM, 2 TB NVMe per node) with PostgreSQL 16.2. Parsing uses one NVIDIA L4 24 GB GPU; reranking and retrieval use 8 NVIDIA A100 80 GB GPUs. Table 2 reports unweighted source-family macro-averages: for metric m and row k, $\bar{m}_k = |\mathcal{F}_k|^{-1} \sum_{f \in \mathcal{F}_k} m_{k,f}$, where $\mathcal{F}_k$ contains the development families represented in that row. The overall drift result uses all 11 development families; the appendix confusion matrix instead pools all 2,000 event predictions. We compute monitoring F1 intervals by 10,000-resample stratified percentile bootstrap with fixed seed 271828, marginal closure intervals by the 95% Wilson method, and closure differences by paired 10,000-resample request bootstrap preserving the five manifest strata. The operational study used de-identified data under local IRB review with waiver of consent and human governance oversight.

3 Results

3.1 Cross-System Historical Case Closure

Every row in Table 1 receives the same 359 prespecified requests and is scored only from its returned case bundle. GuidanceTrace closes all 359 requests, including the 142 negative and 73 hard cases. TemporalSpanQA averages 90.3% $\pm$ 0.7 across five seeds; its seed-1701 output closes 90.8% (326/359). The timestamp-only, bitemporal-database, and RAG+TimeFilter systems close 274, 247, and 151 requests, respectively. One output schema thus separates assertion-store, temporal-database, and document-retrieval designs by their ability to return a reviewer-complete historical answer.

The matched identity intervention isolates a representation mechanism. Holding extraction, scope, provenance fields, span digests, snapshots, indexes, thresholds, hardware, manifest, and scorer fixed leaves task-benchmark F1 at 94.0% for drift and 93.7% for contradiction, while closure falls to 85.5% (307/359). Its 14.5-point paired deficit [10.9, 18.4] concentrates in wrong effective-time versions and incompatible event selection (Appendix Table S6). For the tested selector, independent source-event keys preserve closure information that a shared semantic lookup key removes.

3.2 Failure Stages Distinguish Equal Aggregate Scores

TemporalSpanQA's seed-1701 failures divide into 13 historical-version, 5 answer/bundle, 12 evidence-witness, and 3 reconstruction cases. The two methods that both close 274/359 illustrate why this accounting matters: removing persistent span binding produces 72 evidence failures, while timestamp sorting produces 34 version and 32 reconstruction failures. Because the no-binding adapter hashes selected bytes at submission, its errors isolate revision-to-span retention during selection rather than digest computation. Identical closure totals therefore imply different missing capabilities and different reviewer repair work.

3.3 Grounding, Monitoring, and Source-Family Transfer

On the separate task benchmarks, GuidanceTrace reaches 94.2% drift F1 (95% CI: 93.1–95.2%) and 94.0% contradiction F1 (92.8–95.1%). Negative-class F1 is 96.0% and 94.9%, respectively.

Table 1: Historical case closure on the common 359-request manifest. Deterministic methods report raw counts and 95% Wilson intervals. TemporalSpanQA reports its five-seed mean and standard deviation plus the first listed seed for count-based accounting. Deficits use paired 10,000-resample request bootstraps against GuidanceTrace.

Method	Closure [95% CI]	Deficit [paired 95% CI]
GuidanceTrace	100% (359/359) [98.9–100]	0
Collapsed identity	85.5% (307/359) [81.5–88.8]	14.5 [10.9–18.4]
TemporalSpanQA	90.3% $\pm$ 0.7; seed 1701: 326/359	seed 1701: 9.2 [6.1–12.5]
Baseline D: no persistent binding	76.3% (274/359) [71.7–80.4]	23.7 [19.4–28.2]
Baseline G: no versioning	65.2% (234/359) [60.1–69.9]	34.8 [29.9–39.7]
Baseline A: timestamp-only	76.3% (274/359) [71.7–80.4]	23.7 [19.2–28.4]
Baseline E: bitemporal	68.8% (247/359) [63.8–73.4]	31.2 [26.4–36.0]
Baseline F: RAG+TimeFilter	42.1% (151/359) [37.1–47.2]	57.9 [52.7–63.0]

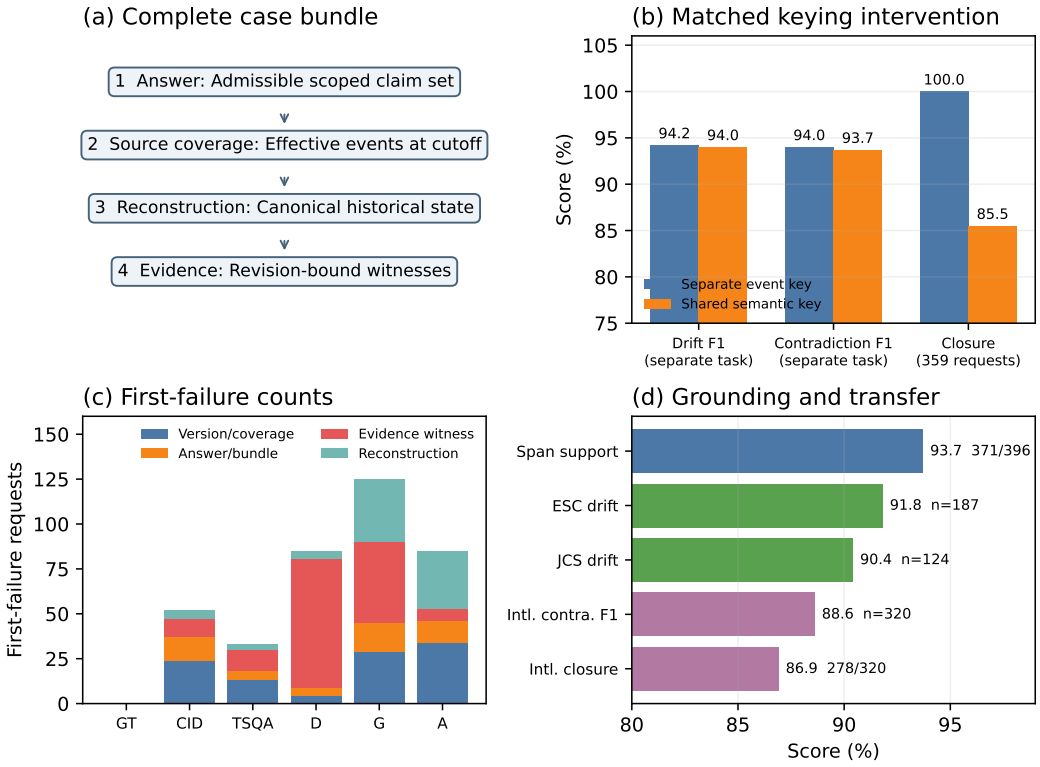


Figure 1: Historical case closure and its evidence. (a) A request closes only when one complete admissible bundle passes all four predicates; negative bundles additionally carry source-coverage certificates. (b) Replacing separately addressable source-event keys with one shared semantic key changes separate-task drift and contradiction F1 by at most 0.3 points but reduces closure on the common manifest by 14.5 points. (c) Exhaustive first-failure stages distinguish equal closure totals; the legend is positioned above the observed counts. (d) Independent grounding and source-chain-isolated transfer; the span-support rate is 371/396, ESC and JCS drift sets contain 187 and 124 events, and the international contradiction set contains 320 ESC/EMA and JCS/PMDA cases.

Expanded-subgroup drift (91.6%) and class-level interaction conflicts (91.7%) are the most difficult subclasses, consistent with scope and qualifier errors. Removing scope alignment reduces drift and contradiction F1 by 12.5 and 14.7 points, and restricting contradiction coverage from 61 to the 20

Table 2: Main drift and contradiction benchmarks. Both tasks include negative classes; class-row point estimates macro-average source-family scores, and values in parentheses are 95% stratified-bootstrap confidence intervals.

Task	Precision	Recall	F1
Guideline drift detection (n=2,000)	96.4	92.1	94.2
	(95.1–97.5)	(90.4–93.7)	(93.1–95.2)
Strengthened recommendations (n=317)	97.1	93.4	95.2
	(95.4–98.4)	(91.0–95.5)	(93.5–96.7)
Weakened recommendations (n=284)	96.3	91.7	93.9
	(94.3–97.9)	(88.8–94.3)	(91.9–95.7)
Reversed recommendations (n=308)	95.2	90.7	92.9
	(93.0–97.0)	(87.8–93.4)	(90.8–94.8)
Expanded subgroup criteria (n=291)	94.8	88.6	91.6
	(92.1–97.0)	(84.8–91.9)	(88.9–94.0)
No drift (negatives, n=800)	97.2	94.8	96.0
	(95.5–98.5)	(92.4–96.8)	(94.3–97.4)
Safety contradiction detection (n=1,200)	97.8	90.6	94.0
	(96.7–98.7)	(88.5–92.5)	(92.8–95.1)
Indication–contraindication (n=247)	98.5	92.3	95.3
	(97.2–99.4)	(89.5–94.8)	(93.5–96.8)
Dose range violations (n=218)	96.7	89.4	92.9
	(94.6–98.3)	(86.2–92.3)	(90.6–95.0)
Class-level interaction conflicts (n=195)	97.1	86.8	91.7
	(95.2–98.5)	(83.3–90.0)	(89.1–94.0)
No contradiction (negatives, n=540)	96.8	93.2	94.9
	(95.1–98.1)	(90.7–95.4)	(93.1–96.5)

most frequent rules costs 6.8 points (Appendix Table S17). End-to-end extraction reaches 86.8% F1 on 500 annotated pairs; this upstream denominator is distinct from the curated closure manifest.

Blinded review finds 371 supported spans among 396 evaluable assertions (93.7%); agreement is $\kappa = 0.84$ on all 400 sampled assertions. Source-chain-held-out drift F1 remains 91.8% on 187 ESC and 90.4% on 124 JCS events. With the 61 contradiction rules and thresholds unchanged, GuidanceTrace records 156 true positives, 16 false positives, 24 false negatives, and 124 true negatives on 320 isolated ESC/EMA and JCS/PMDA cases. These counts yield 90.7% precision, 86.7% recall, and 88.6% F1; 278/320 bundles (86.9%) satisfy complete closure. The two-case difference from the 280 correct contradiction labels is attributable to source-time or evidence-witness failure. This transfer slice shows that the same scorer remains discriminative under new terminology and source chains, while the 18 terminology and 12 scope failures delimit the observed degradation.

3.4 Reviewer Workflow Corroboration

At an independent second site, GuidanceTrace closes 92.2% (166/180) requests under the unchanged rubric, compared with 81.7% (147/180) for TemporalSpanQA and 73.3% (132/180) for the existing workflow. Median review time is 6.8 min, 31.6 min, and 44.7 min, respectively (Appendix Table S10). Two blinded reviewers apply the same case rubric ($\kappa = 0.82$), with third-reviewer adjudication and non-overlapping source chains. The replay connects higher case closure with lower reviewer effort under human governance.

4 Discussion

Historical case closure changes the unit of evaluation from an intermediate prediction to the evidence packet a reviewer can resolve. The 359-request comparison shows that this unit discriminates among assertion stores, temporal databases, and document retrievers under one output schema. Its four predicates also retain information that a scalar success rate discards: the first-failure ledger identifies whether a dated review remains open because of source coverage, bundle construction, witness

validity, or state reconstruction. This diagnostic object is useful even when a system does not adopt GuidanceTrace's storage design.

The mechanism results separate two modeling problems. Scope alignment has the largest measured effect on monitoring, reducing drift and contradiction F1 by 12.5 and 14.7 points when removed. Under fixed extraction and scope, the shared-key intervention instead loses 52 closures, predominantly through version and bundle errors, showing that source-event addressability governs the reviewer handoff for this selector. TemporalSpanQA narrows the architecture gap with bitemporal chunks, learned reranking, and span selection, while the no-binding control isolates the value of retaining revision-to-span links during selection. These complementary findings expose concrete targets for future systems: scope-aware matching, effective-time reranking, set-valued decoding, negative coverage, and revision-bound span selection.

The empirical claim is scoped to reviewer-facing reconstruction over English-language clinical guidance and labels, isolated ESC/EMA and JCS/PMDA source chains, authenticated writes, and two governance teams; patient-outcome effects and trust-root compromise are outside that claim, and qualified reviewers remain the decision authority. Within this regime, 93.7% general-corpus span support, transferred terminology and scope errors, implicit negation, cross-sentence relations, and maintenance of the 61-rule inventory distinguish broad extraction quality from closure on the adjudicated manifest. The international and second-site results nevertheless preserve system separation under new source chains and workflows, providing a concrete basis for extension to additional languages, jurisdictions, and source policies.

5 Conclusion

Historical case closure makes a dated clinical review—the scoped answer, effective source events, reconstructed state, and revision-bound witnesses—a single scored output. The 359-request manifest separates system families from 42.1% to 100% closure, and its exhaustive failure stages identify the exact obligation that keeps each request open. A matched shared-key intervention preserves F1 within 0.3 points on the separate monitoring tasks yet loses 52 closures, directly demonstrating the value of separately addressable source-time events for the tested selector. Source-chain-isolated transfer and independent second-site replay retain the ranking while linking higher closure to lower reviewer effort. HCC therefore supplies both a benchmark target and a diagnostic interface for building learning and retrieval systems whose historical answers arrive complete enough to review.

Ethics Statement

The operational study used de-identified data under local IRB review with waiver of consent and continuous human governance oversight. The system supported evidence review; qualified staff retained every policy and clinical decision.

AI Use Statement

Generative AI tools assisted manuscript preparation, including language-level editing, structural and formatting revisions, and bibliographic consistency checks. The authors reviewed the AI-assisted work and take responsibility for the final submission, including its text, claims, and reported results.

References

Michael Armbrust, Tathagata Das, Liwen Sun, Burak Yavuz, Shixiong Zhu, Mukul Murthy, Joseph Torres, Herman van Hovell, Adrian Ionescu, Alicja Łuszczak, Michał Switakowski, Michał Szafrański, Xiao Li, Takuya Ueshin, Mostafa Mokhtar, Peter Boncz, Ali Ghodsi, Sameer Paranjpye, Pieter Senster, Reynold Xin, and Matei Zaharia. Delta lake: High-performance ACID table storage over cloud object stores. *Proceedings of the VLDB Endowment*, 13(12):3411–3424, 2020. doi: 10.14778/3415478.3415560.

Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010. doi: 10.1136/jamia.2009.002733.

Andrew Bate and Stephen J. W. Evans. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, 18(6):427–436, 2009. doi: 10.1002/pds.1742.

Olivier Bodenreider. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004. doi: 10.1093/nar/gkh061.

Anna Breit, Simon Ott, Asan Agibetov, and Matthias Samwald. Openbiolink: A benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, 36(13):4097–4098, 2020. doi: 10.1093/bioinformatics/btaa274.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023. doi: 10.1038/s41597-023-01960-3.

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. *Scientific Data*, 12(1):1855, 2025. doi: 10.1038/s41597-025-06098-y.

Anton Dignös, Michael H. Böhlen, and Johann Gamper. Temporal alignment. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 433–444. ACM, 2012. doi: 10.1145/2213836.2213886.

Julian H. Elliott, Annie Synnot, Tari Turner, Mark Simmonds, Elie A. Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, David Tovey, Ian Shemilt, and James Thomas. Living systematic review: 1. introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, 91:23–30, 2017. doi: 10.1016/j.jclinepi.2017.08.010.

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

Boris Glavic. Data provenance—origins, applications, algorithms, and models. *Foundations and Trends in Databases*, 9(3–4):209–441, 2021. doi: 10.1561/1900000068.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. doi: 10.1145/3458754.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022. doi: 10.1145/3503161.3548112.

Martin Kaufmann, Amin Amiri Manjili, Panagiotis Vagenas, Peter M. Fischer, Donald Kossmann, Franz Färber, and Norman May. Timeline index: A unified data structure for processing queries on temporal data in SAP HANA. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 1173–1184. ACM, 2013. doi: 10.1145/2463676.2465293.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

Adam Liška, Tomáš Kočiský, 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.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart J. Nelson. Rxnorm: prescription for electronic drug information exchange. *IT Professional*, 7(5):17–23, 2005. doi: 10.1109/MITP.2005.122.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. Biored: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022. doi: 10.1093/bib/bbac282.

Igho J. Onakpoya, Carl J. Heneghan, and Jeffrey K. Aronson. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Medicine*, 14:10, 2016. doi: 10.1186/s12916-016-0553-2.

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

Michiel Overeem, Marten Spoor, Slinger Jansen, and Sjaak Brinkkemper. An empirical characterization of event sourced systems and their schema evolution: Lessons from industry. *Journal of Systems and Software*, 178:110970, 2021. doi: 10.1016/j.jss.2021.110970.

Richard Platt, Jeffrey S. Brown, Melissa Robb, Mark McClellan, Robert Ball, Michael D. Nguyen, and Rachel E. Sherman. The FDA sentinel initiative—an evolving national resource. *New England Journal of Medicine*, 379(22):2091–2093, 2018. doi: 10.1056/NEJMp1809643.

Fotis Psallidas and Eugene Wu. Smoke: Fine-grained lineage at interactive speed. *Proceedings of the VLDB Endowment*, 11(6):719–732, 2018. doi: 10.14778/3199517.3199522.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

S. Trent Rosenbloom, Steven H. Brown, David Froehling, Brent A. Bauer, Dietlind L. Wahner-Roedler, William M. Gregg, and Peter L. Elkin. Using SNOMED CT to represent two interface terminologies. *Journal of the American Medical Informatics Association*, 16(1):81–88, 2009. doi: 10.1197/jamia.M2694. URL https://academic.oup.com/jamia/article-pdf/16/1/81/2574543/16-1-81.pdf.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8017–8026, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.552. URL https://aclanthology.org/2022.acl-long.552/.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018. doi: 10.1016/j.jbi.2017.11.011.

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/.

A Technical Appendix

The appendix details corpus construction, scorer inputs, comparator configuration, manifest composition, first-failure accounting, transfer, extraction slices, and integrity evidence.

A.1 Corpus, annotation, and witness construction

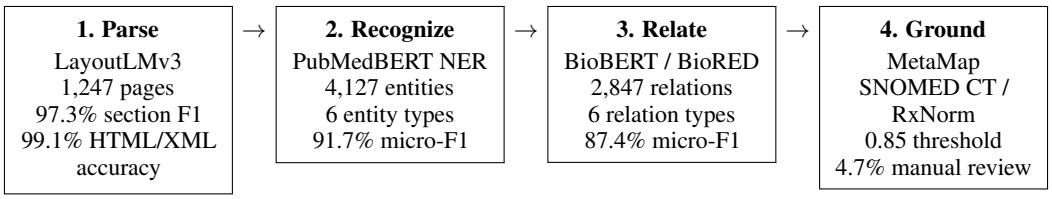
The source inventory spans longitudinal guideline, label, and operational streams. Benchmark sampling preserves substantial negative classes and hard strata, and the witness schema retains the exact fields consumed by the case-closure scorer.

Table S1: Time-versioned source streams used to construct the evaluation corpus and pilot context.

Source	Type	Count	Years	Content scope
Specialty society guidelines	Unstructured	3,842 versions	2014–2025	Recommendations, contraindications, dosing guidance, subgroup criteria
Regulatory drug labels	Semi-structured	12,417 updates	2006–2025	Indications, warnings, drug–drug interactions, boxed warnings, dose adjustments
De-identified EHR events	Structured	4.8M events	2018–2024	Medication administrations, laboratory results, problem lists (event-level timestamps)

Table S2: Benchmark construction and annotation protocol for the drift and contradiction tasks.

Attribute	Drift benchmark / Contradiction benchmark
Sample size	2,000 events / 1,200 cases
Positive rate	60% drift / 55% contradiction
Guideline families / contradiction strata	11 development families (ESC/JCS held out) / 9 source-pair strata
Annotator 1	MD, board-certified internal medicine, 12 years experience
Annotator 2	MD, board-certified internal medicine, 8 years experience
Pharmacist reviewer	PharmD, 15 years clinical experience (contradiction only)
Annotation rubric	Drift class (4 categories) + scope tuple + evidence spans / Contradiction type (3 categories) + resolution status
Initial agreement	91.3% / 88.7%
Cohen's κ	0.87 / 0.84
Adjudication	Consensus discussion for all disagreements
Sampling strategy	Stratified by specialty, drift class, version gap; 40% negative / Stratified by source pair, rule category; 45% negative
Difficulty bins	Easy (clear wording change), Medium (implicit), Hard (multi-hop reasoning): 35%/45%/20% / 40%/40%/20%



Training adds BC5CDR, ChemProt, 2,841 guideline sentences, BioRED, and 3,512 guideline relations. On one NVIDIA L4 (24 GB; batch size 32), the pipeline processes 847 assertions/minute and reaches 86.8% end-to-end assertion F1.

Figure S1: The assertion path preserves a visible progression from document structure to ontology-normalized relations. Held-out counts and stage scores are shown inside each box; the resulting scoped assertions feed every shared-extraction comparator.

Table S3: Canonical scorer-facing object, event, and witness fields. Submitted arrays use the ordering and null rules in Sec. 2.1.

Field	Type	Description
`{H,E,W,Z,R}`	JSON object	Exact top-level claim, event, witness, state, and coverage members
`R.coverage_type`	ENUM	`POSITIVE`, `NO_CHANGE`, or `NO_CONFLICT`
`R.entries`	Sorted array	Empty for positive cases; revision–claim-digest pairs for `NO_CHANGE`; source–scoped-claim pairs for `NO_CONFLICT`
`claim_id`	CHAR(64)	SHA-256 of (subject, predicate, object, scope); enables cross-source dedup
`assertion_id`	CHAR(64)	SHA-256 of claim, source, source version, and effective interval
`subject`	VARCHAR(64)	Subject concept CUI (SNOMED CT / RxNorm)
`predicate`	VARCHAR(64)	Relationship type (INDICATES, CONTRAINDICATES, etc.)
`object`	TEXT	Object concept CUI or literal value
`scope`	JSONB	Scope tuple (condition, stage, line, age_band, comorbidities)
`source_id`	VARCHAR(128)	Originating guideline/label/EHR identifier
`published_at`	TIMESTAMPTZ	Publication or encounter time; retained as provenance
`valid_from/to`	TIMESTAMPTZ	Half-open effective interval $[t_{\mathrm{from}}, t_{\mathrm{to}})$
`revision_digest`	CHAR(64)	SHA-256 commitment to the frozen source-revision bytes
`offset_start/end`	BIGINT	Half-open character interval in normalized revision text
`span_digest`	CHAR(64)	SHA-256 of normalized bytes at the submitted character interval
`confidence`	REAL	Rule-derived confidence [0,1]
`source_version`	VARCHAR(64)	Immutable version identifier within the registered source
`status`	VARCHAR(16)	`active`, `superseded`, or `withdrawn` tombstone

The source inventory and task benchmark provide the upstream context for the closure manifest. The schema and pipeline then identify the exact event fields consumed by the scorer; in particular, effective intervals and revision-bound witnesses remain distinct from publication timestamps and semantic claim identity.

A.2 Evaluation-object comparison

Table S4 locates the benchmark among named neighboring evaluations. The marks refer to native scoring requirements, not to whether a method could be adapted to emit additional fields.

Table S4: Native scoring requirements in neighboring evaluation families. A check denotes a required field, a tilde partial or adapter-level support, and a dash an objective outside the standard scoring unit; the final column identifies whether the fields are scored as one jointly closing object.

Evaluation family	Semantic alignment	Source-time event	Point-in-time state	Span check	Frozen manifest	Joint closure
Biomedical KG / CDS Chandak et al. (2023); Breit et al. (2020)	✓	∼	–	∼	–	–
Situated/streaming QA Zhang & Choi (2021); Liška et al. (2022)	∼	∼	✓	–	∼	–
Temporal RAG Ouyang et al. (2025); Chen et al. (2025)	∼	∼	✓	∼	∼	–
Citation attribution Gao et al. (2023); Rashkin et al. (2023)	∼	–	–	✓	∼	–
Temporal database Dignös et al. (2012); Kaufmann et al. (2013)	–	✓	✓	∼	–	–
Living guidance Elliott et al. (2017)	✓	✓	∼	∼	–	–
Historical case closure (ours)	✓	✓	✓	✓	✓	✓

A.3 Frozen manifest and first-failure accounting

The manifest includes positive and negative requests across source pairs, cutoff ages, and difficulty levels. Because GuidanceTrace closes all 359 requests, its closure is 100% in every listed stratum; the composition table makes the support of that statement explicit.

Table S5: Composition of the 359-request historical closure manifest. Categories are mutually exclusive within each block and every block reconciles to 359. The manifest was stratified and frozen before comparator outputs were scored.

Block	Category	N	Category	N	Category	N
Request type	Superseded guidance	101	Cross-source conflict	92	Scope qualifier	88
	No-change verification	78				
Source pair	Guideline–guideline	143	Label–label	97	Guideline–label	119
Cutoff age	<6 months	85	6–24 months	111	2–5 years	96
	>5 years	67				
Case status	Positive	217	Negative	142		
Difficulty	Easy	126	Moderate	160	Hard	73

Aggregate closure is accompanied by one earliest reviewer-blocking stage per failed request. Table S6 applies the exhaustive mapping in Sec. 2.1, so each non-closing request contributes exactly once.

Table S6: Mutually exclusive first-failure counts on the frozen manifest. Version includes effective-time and negative coverage; answer/bundle includes malformed, missing, non-terminating, semantic, and incompatible-set outputs; evidence includes incomplete, non-unique, or unverifiable witnesses. Row failures equal $359-$closures in Table 1.

Method	Closed	Version / coverage	Answer / bundle	Evidence witness	Reconstruction
GuidanceTrace	359	0	0	0	0
Collapsed identity	307	24	13	10	5
TemporalSpanQA	326	13	5	12	3
Baseline D: no persistent binding	274	4	5	72	4
Baseline G: no versioning	234	29	16	45	35
Baseline A: timestamp-only	274	34	12	7	32
Baseline E: bitemporal	247	26	16	21	49
Baseline F: RAG+TimeFilter	151	67	44	61	36

A.4 Comparator configuration and family transfer

Table S7 reports separate task-benchmark F1 and rollback context without treating those values as scores on the closure manifest.

Table S7: Intervals on the separate drift and contradiction task benchmarks, plus rollback latency. These F1 values are system context and are not measured on the 359 closure requests. F1 intervals use 10,000 stratified case bootstraps.

Method	Drift F1 [95% CI]	Contra. F1 [95% CI]	Rollback p95
GuidanceTrace	94.2 (93.1–95.2)	94.0 (92.8–95.1)	5.13 s
Collapsed identity	94.0% (92.8–95.1)	93.7% (92.4–94.9)	5.21 s
TemporalSpanQA	92.0% (90.5–93.4)	90.8% (89.0–92.4)	–
Baseline D: no persistent binding	94.1 (92.9–95.2)	93.8 (92.5–94.9)	5.29 s
Baseline G: no versioning	91.4 (89.8–92.9)	90.2 (88.4–91.9)	–
Baseline A: timestamp-only	88.1 (86.4–89.7)	86.7 (84.8–88.5)	61.37 s
Baseline E: bitemporal	89.3 (87.6–90.9)	87.8 (85.9–89.6)	8.72 s
Baseline F: RAG+TimeFilter	72.4 (69.8–74.9)	68.3 (65.2–71.3)	–

Table S8 records what each method indexes, selects, and returns. Shared inputs isolate representation choices, while document comparators preserve a separate retrieval architecture.

Table S8: Comparator configuration frozen before test scoring. Shared entries mean that the identical GuidanceTrace extraction and scope outputs are consumed; document entries use the same eligible source documents and cutoff timestamps.

Method	Input / temporal index	Selection and output	Tuning, randomness, and shared components
GuidanceTrace	Assertions / snapshots	Active event, stored span	Development families; deterministic; full extraction and scope
Collapsed identity	Assertions / snapshots	Collapsed key, stored span	Same thresholds; deterministic; all components shared except keying
TemporalSpanQA	Documents / bitemporal chunks	Top-20 reranked chunks, constrained bundle	Development families; five seeds; eligible documents and cutoffs shared
Baseline D	Assertions / snapshots	Active event, selected bytes; adapter digest	Same thresholds; no persistent revision–span binding
Baseline G	Assertions / current-state index	Current event, extracted span	Same thresholds; deterministic; extraction and scope shared
Baseline A	Assertions / timestamp sort	Latest pre-cutoff record, source text	Development families; deterministic; extracted records shared
Baseline E	Assertions / PostgreSQL bitemporal	SQL valid-time result, source text	Development families; deterministic; extracted records shared
Baseline F	Documents / Chroma+time filter	Top-k, GPT-4-turbo bundle	$k \in \{5, 10, 20, 40\}$; temperature 0; seed 1701 reported

TemporalSpanQA obtains 90.3% $\pm$ 0.7 mean closure across all five listed seeds, 93.6% historical-version accuracy, and 94.2% unique-span validity. Seed 1701, the first listed seed rather than the best-scoring seed, supplies the raw 326/359 count and first-failure row. Median/p95 request latency is 0.74 s/1.86 s; indexing and reranking require 11.6 GPU-h, and constrained decoding uses temperature 0.

The international and second-site tables test the scorer under non-overlapping source chains. Their raw counts expose the distinction between a correct contradiction label and a complete historical bundle.

Table S9: Source-chain-isolated international contradiction transfer with the 61-rule package and thresholds frozen. Counts are TP/FP/FN/TN over 180 positive and 140 negative cases. Closure requires all four scorer predicates, so it remains stricter than contradiction F1.

Method	TP	FP	FN	TN	Precision	Recall	F1 / closure
GuidanceTrace	156	16	24	124	90.7%	86.7%	88.6% / 86.9% (278/320)
Collapsed identity	150	19	30	121	88.8%	83.3%	86.0% / 79.1% (253/320)
TemporalSpanQA	148	18	32	122	89.2%	82.2%	85.5% / 82.2% (263/320)

GuidanceTrace closure by source pair is 89.7% (165/184) on ESC/EMA and 83.1% (113/136) on JCS/PMDA (aggregate Wilson 95% CI: 82.7–90.1). The principal closure losses are terminology transfer (18 cases) and scope transfer (12 cases).

Table S10: Independent second-site retrospective replay on 180 requests frozen before scoring. All methods use the same all-or-nothing rubric; intervals are 95% Wilson intervals. Reviewer time measures governance effort per request.

Method	Closure [95% CI]	Version / span validity	Review time median / p95
GuidanceTrace	92.2% (166/180) [87.4–95.3]	96.7% (174/180) / 95.0% (171/180)	6.8 min / 18.4 min
TemporalSpanQA	81.7% (147/180) [75.4–86.6]	91.1% (164/180) / 87.2% (157/180)	31.6 min / 78.2 min
Existing workflow	73.3% (132/180) [66.4–79.3]	86.1% (155/180) / 82.8% (149/180)	44.7 min / 112.5 min

Two blinded reviewers independently applied the rubric ($\kappa = 0.82$); a third reviewer adjudicated disagreements. Source-version chains do not overlap the development corpus.

A.5 Comparator capabilities and component diagnostics

The capability matrix distinguishes emitted scorer support from measured closure. Extraction and monitoring slices then localize residual errors to cross-sentence relations, implicit negation, multi-hop class reasoning, and scope normalization.

Table S11: Scorer-predicate capability matrix for systems measured on the frozen manifest. A check denotes native support, a tilde denotes adapter-provided support, and a dash denotes no emitted support. Every row emits the typed schema and is scored on joint closure; capability does not imply a successful measured case, whose outcomes appear in Table 1.

Method	Semantic alignment	Source-time event	Point-in-time state	Span check	Typed output	Scored jointly
GuidanceTrace	✓	✓	✓	✓	✓	✓
Collapsed identity	✓	∼	✓	✓	✓	✓
TemporalSpanQA	✓	∼	✓	✓	✓	✓
Baseline A: timestamp-only	∼	∼	∼	∼	✓	✓
Baseline D: no persistent binding	✓	✓	✓	∼	✓	✓
Baseline E: bitemporal	∼	✓	✓	∼	✓	✓
Baseline F: RAG+TimeFilter	∼	∼	✓	∼	✓	✓
Baseline G: no versioning	✓	∼	∼	✓	✓	✓

Table S12: Held-out extraction performance by pipeline stage. The last row reports end-to-end assertion extraction.

Component	Precision	Recall	F1 / Accuracy
Document parsing (section)	98.1 (96.8–99.1)	96.5 (94.9–97.8)	97.3
Named entity recognition	93.2 (91.5–94.7)	90.3 (88.4–92.1)	91.7
Relation extraction	89.6 (87.4–91.5)	85.3 (82.9–87.5)	87.4
Ontology grounding	–	–	95.3 (acc)
End-to-end (assertion)	89.2 (86.8–91.4)	84.6 (81.7–87.3)	86.8

Table S13: End-to-end extraction performance by source type. Values in parentheses are 95% bootstrap confidence intervals.

Source Type	N (pairs)	Precision	Recall	F1
PDF guidelines	312	87.4 (84.1–90.3)	82.1 (78.4–85.5)	84.7 (81.9–87.3)
HTML/XML labels	188	92.7 (89.4–95.3)	89.8 (86.1–92.9)	91.2 (88.7–93.4)
Overall	500	89.2 (86.8–91.4)	84.6 (81.7–87.3)	86.8 (84.7–88.8)

Native schema support does not determine closure by itself: every comparator emits the scored object, yet the main comparison spans 42.1–100%. The extraction tables identify a separate upstream bottleneck, with relation extraction and PDF guideline parsing contributing most of the end-to-end loss.

Table S14: Drift transfer across within-portfolio and source-chain-isolated society holdouts. Within-portfolio rows report 95% stratified-bootstrap intervals; the isolated ESC/JCS rows report fixed test-set sizes and F1, visualized in Fig. S3.

Held-out Society	Precision	Recall	F1
ACC/AHA (cardiology)	95.1 (92.8–96.9)	89.8 (86.7–92.5)	92.4 (90.3–94.2)
NCCN (oncology)	96.3 (94.2–97.9)	91.4 (88.5–93.9)	93.8 (91.8–95.5)
Development benchmark (11 families)	96.4 (95.1–97.5)	92.1 (90.4–93.7)	94.2 (93.1–95.2)

Source-chain-isolated external holdout	N	F1
ESC	187	91.8%
JCS	124	90.4%

Table S15: Drift-detection failure envelope on hard subsets. Delta is measured against the overall drift F1 of 94.2%.

Hard Subset	N	F1	Δ vs 94.2%	Primary Error
Complex scope ($\geq$3 attr)	287	88.4%	–5.8 pp	Alignment ambiguity
Cross-sentence relations	412	85.1%	–9.1 pp	Context truncation
Implicit negation	89	79.8%	–14.4 pp	Missing trigger words
Multi-hop class reasoning	156	82.5%	–11.7 pp	Class hierarchy gaps
Non-English terminology	73	86.7%	–7.5 pp	Vocabulary mismatch

Table S16: Scope-alignment coverage by source type. Excluded assertions are sent to manual review.

Source Type	Full Scope	Partial	Excluded	Total (M)
Guidelines	87.3%	8.5%	4.2%	31.47
Labels	82.1%	11.1%	6.8%	28.62
EHR events	71.8%	16.1%	12.1%	18.34
Source-count-weighted overall	81.8%	11.2%	7.0%	78.43

Full scope: all 5 fields populated. Partial: 1–2 wildcards. Excluded: $\geq$3 missing fields.

The source-family and hard-subset slices connect the transfer errors to concrete language phenomena. Scope is the common mechanism: complex qualifiers lower drift F1, 7.0% of assertions lack enough populated scope fields for automated alignment, and the matched no-scope intervention produces the largest monitoring loss.

The ablations separate semantic monitoring from temporal reconstruction. Scope alignment and rule coverage change detection quality, whereas snapshot cadence leaves F1 fixed and changes rollback latency; the pooled confusion matrix localizes the remaining semantic confusions.

Table S17: Ablations with unchanged task label spaces for scope alignment, contradiction-rule coverage, and snapshot cadence.

Configuration	Drift F1	Δ	Contra. F1	Δ	Rollback p95	Δ
GuidanceTrace (full system)	94.2%	—	94.0%	—	5.1 s	—
Ablation 1: No scope alignment	81.7%	–12.5 pp	79.3%	–14.7 pp	5.1 s	—
Ablation 2: Top-20 rules only (not 61)	94.2%	—	87.2%	–6.8 pp	5.1 s	—
Ablation 3: Daily snapshots (not per-batch)	94.2%	—	94.0%	—	18.7 s	+13.6 s

pp: percentage points. Rollback p95 measured at 78M assertions.

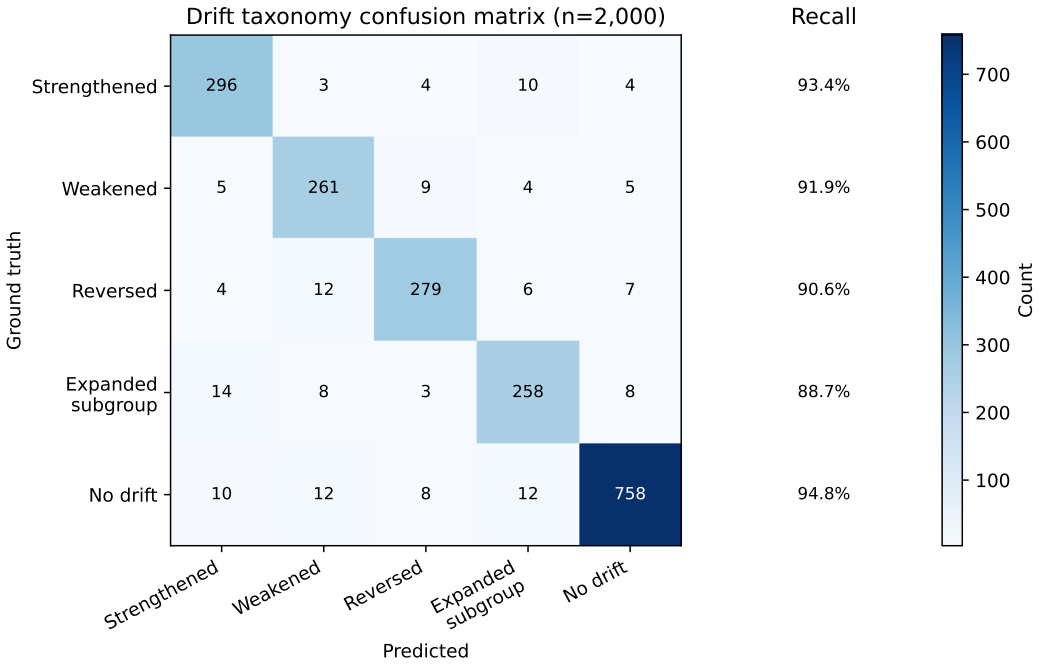


Figure S2: Drift taxonomy confusion matrix for the guideline drift benchmark (n=2,000 events). Rows indicate ground-truth classes and columns model predictions; row totals match the main benchmark breakdown: 317 strengthened, 284 weakened, 308 reversed, 291 expanded-subgroup, and 800 no-drift events. Off-diagonal mass is concentrated in adjacent semantic classes, including expanded subgroup → strengthened (4.8%) and weakened → reversed (3.2%). Counts and right-hand recalls are pooled over events; Table 2 instead reports the unweighted source-family macro-average defined in Sec. 2.4, producing class-recall differences of at most 0.2 points. The two views expose errors at semantic boundaries and in scope alignment without conflating their averaging rules.

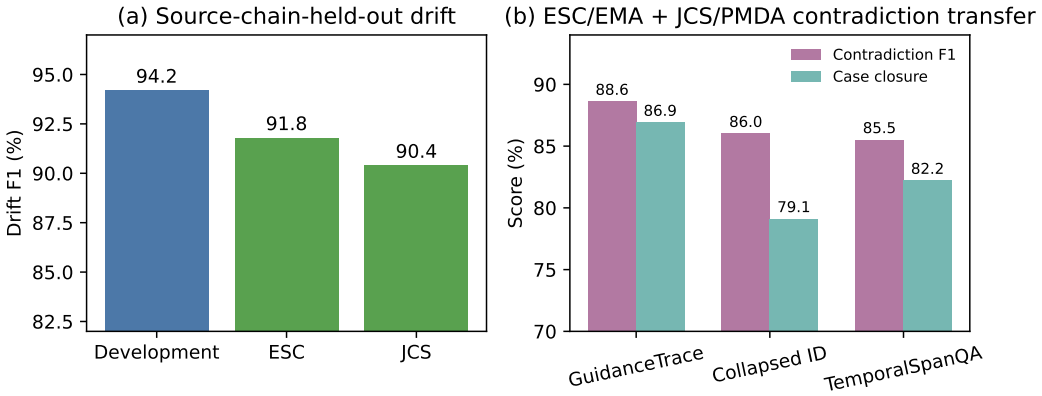


Figure S3: Held-out transfer under complete source-version-chain isolation. (a) Drift F1 on ESC and JCS remains above 90% without development-family overlap; exact sample sizes and F1 values appear in Table S14. (b) With the 61-rule package and all thresholds unchanged, GuidanceTrace, the collapsed-ID control, and TemporalSpanQA retain 88.6%, 86.0%, and 85.5% contradiction F1 on 320 ESC/EMA and JCS/PMDA cases. Joint closure separates the systems further by requiring source-time and span validity; raw confusion counts and source-pair slices appear in Table S9.

The held-out source chains preserve the ordering between GuidanceTrace, collapsed identity, and TemporalSpanQA while exposing terminology and scope as the dominant transfer costs. Together, these displays show why the case-level failure ledger is needed: systems can expose the same fields while differing in event selection, evidence binding, and reconstructed state.

A.6 Integrity and operational evidence

The matrix ties each protected object and attack capability to a measured comparator, overhead, failure mode, and evidence trace. The original and adaptive simulations test post-registration tamper evidence within the authenticated-write, trusted-operator boundary; the workflow table is descriptive operational context.

Table S18: Tamper-evidence matrix under authenticated writes, a trusted operator, and an immutable revision registry. SHA-256 supplies content commitments; authenticity derives from the write and registry boundary, not from a keyed hash.

Object and attack	Boundary / property	Comparator / adaptive evidence	Utility, residual / trace
Evidence span; normalized-text edit	SHA-256 commitment / tamper evidence	Full binding: 1,172/1,200 (97.7%); no-binding D: 842/1,200	0.36 ms adaptive p95; normalization-equivalent edits / Table S19
Effective interval; backdate	Assertion digest plus authenticated append / temporal consistency	Timestamp-only A; precision and boundary probes	5.13 s rollback p95; sub-second precision loss / Tables 1, S19
Source ID; substitution	Registry integrity / event attribution	Collapsed identity; isolated source chains	94.2% drift F1; cross-source aliasing / Tables 1, S9
Predicate; relation edit	Rerunnable extraction / assertion integrity	No-binding D; adaptive relation edits	94.0% contradiction F1; ontology-equivalent wording / Table S19
Active record; deletion	Append-only log / completeness evidence	Bitemporal E; held-out deletion timing	3/50,000 false positives; snapshot-boundary race / Table S19
Snapshot; prior-state rewrite	Pointer and key integrity / reconstruction	No-versioning G; adaptive boundary probe	26,917 assertions/s; boundary ordering / Tables 1, S19

Across the original in-scope simulation, GuidanceTrace detects 5,791/5,800 attacks. Removing persistent revision–span binding detects 842/1,200 adaptive attacks, isolating that binding within the authenticated-write boundary.

The adaptive comparison isolates persistent revision–span binding: detection falls from 1,172/1,200 to 842/1,200 when that binding is removed, while the attack definitions keep the authenticated registry fixed.

Table S19: Original in-scope and held-out adaptive tamper evaluations by attack type, detection mechanism, and miss profile.

Attack Type	N	Detected	TPR	Detection Mechanism
Evidence span modification	2,000	1,994	99.7%	`span_digest` mismatch
Effective-time backdating	1,500	1,498	99.9%	Effective-interval component of `assertion_id` plus append-order check
Source ID substitution	1,000	1,000	100.0%	`assertion_id` hash
Predicate alteration	800	800	100.0%	`assertion_id` hash
Silent deletion	500	499	99.8%	Snapshot diff
Total	5,800	5,791	99.84%	–

Missed detections: 6 whitespace-only changes below the normalization threshold, 2 sub-second timestamp precision losses, and 1 snapshot-boundary race.

False positives: 3 of 50,000 unmodified assertions (0.006%; 95% CI: 0.001–0.018%).

Original-set detection latency: 0.12 ms (p50) and 0.31 ms (p95) per assertion.

Held-out adaptive set: GuidanceTrace detects 1,172/1,200 attacks (97.7%), compared with 842/1,200 without persistent revision–span binding; adaptive p95 latency is 0.36 ms per assertion.

The governance pilot records the same reviewer-facing quantities over sustained operation: actionable changes, detection and triage time, policy-update time, alert precision, and staff effort.

Table S20: Six-month single-site governance pilot reporting descriptive workflow measures under human oversight.

Metric	Pilot (GuidanceTrace)	Manual workflow
Actionable changes surfaced	127 over 6 months	23 over 6 months
Time-to-detection (median)	0.8 days	47 days
Time-to-triage, high priority (median)	4.2 hours	18.3 days
Time-to-policy-update (median)	8.7 days	34.2 days
False positive rate (drift alerts)	4.1%	N/A
False positive rate (contradiction alerts)	2.4%	N/A
Staff time per detected change	6.3 min	4.2 hours
Inter-rater reliability (Fleiss' κ)	0.84	–

Together, the integrity probes and governance pilot connect the scorer-facing representation to two reviewer-relevant properties: registered evidence remains traceable under the tested edits, and the resulting bundles remain usable in sustained human review.