DISENTANGLING UNCERTAINTY FROM EVIDENCE VALIDITY IN CLINICAL TRIAL SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective clinical trial screening has two independently observable events: whether an eligibility label is decisive and whether its cited evidence was available, patient-bound, and intact at the protocol cutoff. ACTAUDIT represents these events separately: automatic action requires both a singleton split-conformal set and deterministic source, time, patient, span, and integrity checks over a nonempty citation set. To identify the contribution of each condition, the evaluation fixes every retrieved-evidence and label-score tuple and changes only the routing policy. On 21,073 grouped-holdout patient–criterion pairs, nearly equal-coverage policies (67.2% versus 67.3%) yield 0.40% versus 1.27% false-eligible (FE) risk and 0 versus 11 post-cutoff citations. Removing evidence validity from the identical conformal threshold adds 3.4 coverage points, 0.31 FE points, and 14 post-cutoff citations; removing singleton uncertainty adds 12.3 coverage points and 1.31 FE points. Across five prespecified operating points, the joint rule retains at least a 0.65-point nearest-coverage FE advantage over scalar confidence, and frozen transfer to 6,240 FHIR-native pairs reproduces the separation (0.61% versus 1.68% FE). Evidence removal also changes the label or forces deferral for 88.3% of audited note-dependent actions. The results establish structural evidence validity as an independently measurable determinant of selective action in the evaluated clinical cohorts.

1 INTRODUCTION

For binary selective prediction, a singleton split-conformal set is a threshold event on the winning-label score. That event is unchanged if a pre-cutoff laboratory citation is replaced by a later record, a different patient's record, or an unverifiable note span while the label score stays fixed. Clinical trial screening exposes the consequence: uncertainty can determine whether a label is decisive, but not whether its evidence was available and correctly bound at the screening time. False-eligible (FE) actions can initiate inappropriate outreach or propagate a protocol error, while deferral preserves coordinator review at a measurable workload cost (Getz & Campo, 2017; Fogel, 2018; Ross et al., 2010).

This paper turns that invariance into a paired routing intervention. Each patient–criterion example contributes one frozen tuple containing the predicted label, normalized class scores, conformal set, and cited evidence. ACTAUDIT routes the label to automatic action only when the set is singleton and every citation passes a structural predicate over source identity, temporal availability, patient binding, span offsets, and content integrity. The paired comparison changes this predicate while preserving retrieval, reranking, context, adjudication, decoding, and calibration data. It therefore measures whether evidence validity changes the action set after the predictor has already spoken.

The primary holdout makes this effect visible at matched coverage. ACTAUDIT acts on 67.2% of 21,073 pairs at 0.40% FE with no post-t_0 citation among 129,844 cited objects; the nearest scalar-confidence point acts on 67.3% at 1.27% FE and admits 11 post-t_0 citations. At the identical conformal threshold, removing only the evidence predicate adds 3.4 coverage points, 0.31 FE points, and 14 temporal violations. Removing only singleton uncertainty instead adds 12.3 coverage points and 1.31 FE points. These two one-factor interventions distinguish evidentiary selection from uncertainty selection without attributing either effect to a different model.

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.

Recent work brings conformal prediction to RAG correctness, generation risk, factuality, and output selection (Li et al., 2024; Kang et al., 2024; Feng et al., 2025; Gui et al., 2024). Clinical trial matching has also studied confidence-based abstention (León Tramontini et al., 2026). Those advances motivate selective action, but do not measure the paired effect of a deterministic per-citation validity predicate on directional clinical error. Evidence-grading frameworks instead make support sufficiency an explicit condition for clinical claims (Lin et al., 2026). ACTAUDIT complements that direction by isolating structural provenance and integrity before action and auditing semantic support separately.

ACTAUDIT's technical contribution is an identifiable decomposition of selective clinical action. Its evaluation object records acted-on coverage, directional FE and false-ineligible (FI) risk, and action-level structural evidence failures. Matched-stack closure makes the evidence intervention exact, while the reverse removal quantifies the role of singleton uncertainty. The 84,291-pair, 86-site study then tests whether this decomposition persists across a five-point operating neighborhood, an untouched FHIR-native service, unseen-site and temporal shifts, semantic evidence judgments, and integrity attacks. Together, these experiments show that evidence validity changes which frozen prediction tuples receive automatic action and locate the mechanisms responsible for that change.

2 RELATED WORK

Clinical trial matching and abstention. Eligibility systems span computable phenotypes, rule-based pre-screening, learned retrieval and ranking, criterion-level reasoning, and fairness-aware matching (Kang et al., 2017; Miotto & Weng, 2015; Ni et al., 2015; Shivade et al., 2014; Hripcsak & Albers, 2013; Jin et al., 2024; Gupta et al., 2024; Nievas et al., 2024; Alkhoury et al., 2025; Yuan et al., 2023; Chang et al., 2023). The closest task-specific abstention study evaluates verbalized confidence and optional refusal in a patient–trial RAG pipeline (León Tramontini et al., 2026). ACTAUDIT instead uses a calibration-derived singleton event, directional clinical losses, and an independently testable evidence predicate. Its paired intervention asks which already-produced criterion labels can enter automatic action, rather than improving the underlying matcher.

Conformal selection for foundation models and RAG. Conformal prediction, selective classification, and risk control formalize prediction sets and coverage–risk trade-offs (Vovk et al., 2005; Angelopoulos & Bates, 2021; Geifman & El-Yaniv, 2019; Romano et al., 2020; Angelopoulos et al., 2024; Qiu et al., 2023). TRAQ constructs retrieval and answer sets with correctness guarantees; C-RAG bounds generation risk; Conformal-RAG targets sub-claim factuality; and Conformal Alignment selects outputs against a learned human-alignment criterion (Li et al., 2024; Kang et al., 2024; Feng et al., 2025; Gui et al., 2024). These methods make uncertainty or predicted quality a selection variable. ACTAUDIT studies an orthogonal intervention: whether a source-, patient-, time-, span-, and hash-valid citation set changes routing when the conformal event and all prediction tuples are held fixed. Directional FE is measured separately because marginal conformal coverage does not encode the asymmetric screening loss.

Evidence grounding and clinical claim gates. RAG and attribution methods improve knowledge access, citation generation, correction, grounding, and factual consistency (Lewis et al., 2020; Gao et al., 2023c;a;b; Aly et al., 2024; Huang et al., 2024). Evidence-graded decision authorization additionally conditions clinical claims on a rule-based sufficiency level (Lin et al., 2026). ACTAUDIT separates structural and semantic evidence questions on the same actions. A deterministic pre-action predicate verifies identity, temporal availability, patient binding, offsets, and integrity; blinded judgments and counterfactual removal then test support and decision relevance. The matched routing intervention, rather than the existence of an evidence gate alone, is the distinctive object evaluated here.

3 METHOD: EVIDENCE-CONDITIONED SELECTIVE ACTION

For a policy g locked before evaluation on dataset D, let $A_g = \{x : g(x) \neq \text{DEFER}\}$, $\widehat{\text{Cov}}_D(g) = |A_g|/|D|$, and $\widehat{R}_{\text{FE},D}(g)$ be FE risk over acted-on pairs. With structural evidence predicate V, the

empirical target region is

$$\widehat{\mathcal{A}}_{\epsilon,\gamma}(D) = \big\{g \in \mathcal{G} : \widehat{R}_{\mathrm{FE},D}(g) \le \epsilon,\ \widehat{\mathrm{Cov}}_D(g) \ge \gamma, \quad V(E(x)) = 1\ \ \forall x \in A_g\big\}. \tag{1}$$

We prespecify $\epsilon = 1.0\%$ and monitored coverage $\gamma = 50\%$. Equation 1 evaluates locked policies; it does not select a holdout-optimal operating point. The hat makes membership cohort-indexed, and the reported coordinate vector and intervals retain conformal coverage, directional FE, acted-on coverage, and structural validity as distinct quantities.

Paired routing intervention. Let $F_\theta(x)$ return the shared tuple $(\hat{y}, \hat{\pi}, E, C)$. Matched-stack closure fixes θ, retrieved candidates, reranking, context budget, decoding, and the calibration partition for every $g \in \mathcal{G}$. The two identifying contrasts evaluate the conjunction against (i) the identical singleton event without $V(E)$ and (ii) the identical $V(E)$ event without singleton uncertainty. Scalar confidence and an FE-budgeted directional score rule place these one-factor interventions relative to conventional score routing. Every policy receives the same tuple.

Conjunctive routing. ACTAUDIT maps each atomic criterion and time-indexed chart x to $\hat{y} \in \{\text{MET}, \text{NOT-MET}\}$, normalized class scores, and citation set $E(x)$. Routing joins independently testable uncertainty and evidence conditions:

$$g(x) = \begin{cases} \hat{y}, & |C(x)| = 1 \ \wedge\ V(E(x)) = 1, \\ \text{DEFER}, & \text{otherwise}, \end{cases} \tag{2}$$

where $C(x)$ is the conformal prediction set and V is the evidence-validity predicate. With scores $s_i = 1 - \hat{\pi}(y_i \mid x_i)$ and calibration quantile $\hat{q}_\alpha$ at empirical level $\lceil (1-\alpha)(n+1) \rceil / n$,

$$C(x) = \{y : \hat{\pi}(y \mid x) \ge 1 - \hat{q}_\alpha\}. \tag{3}$$

Split conformal uses the class scores for ranking and does not require them to be calibrated probabilities. The primary $\alpha = 0.03$ is a prespecified design point fixed before holdout scoring, as are all comparator thresholds. Score ties retain both labels and therefore defer. The neighboring α values form a prespecified operating-point study; their holdout coordinates characterize sensitivity rather than select or revise the primary policy.

Binary-threshold equivalence and controls. For two normalized class scores, $|C(x)| = 1$ is equivalent to a calibration-derived threshold on the winning class, apart from ties. Conformal-only is therefore the exact threshold control for the uncertainty condition used by ACTAUDIT; validity-only supplies the reverse component removal. A separately tuned global-confidence rule tests conventional scalar routing, while the FE-budgeted comparator learns distinct inclusion- and exclusion-direction thresholds on calibration by maximizing coverage subject to a one-sided 95% FE bound of 1.0%. The component claim rests on the exact one-factor contrasts, while the broader score controls characterize the surrounding risk–coverage behavior.

Audit estimands. Let $p(x) \in \{\text{INCLUSION}, \text{EXCLUSION}\}$ denote criterion polarity and y the resolved gold label. Directional FE loss is $\ell_{\mathrm{FE}}(x) = \mathbb{I}[p(x) = \text{INCLUSION}, \hat{y} = \text{MET}, y = \text{NOT-MET}] + \mathbb{I}[p(x) = \text{EXCLUSION}, \hat{y} = \text{NOT-MET}, y = \text{MET}]$, with FI defined by reversing the predicted and gold states. For each policy, $\widehat{R}_{\mathrm{FE}} = |A_g|^{-1} \sum_{x \in A_g} \ell_{\mathrm{FE}}(x)$ and $\widehat{R}_V = 1 - |A_g|^{-1} \sum_{x \in A_g} V(E(x))$ use the acted-on set, while coverage uses all scored pairs. Reporting the coordinate vector $(\mathrm{Cov}, \widehat{R}_{\mathrm{FE}}, \widehat{R}_{\mathrm{FI}}, \widehat{R}_V)$ prevents a coverage increase from concealing directional or evidentiary cost. Because policies act on the same examples and shared tuples, policy contrasts are paired; trial- and site-stratified resampling preserves the dependence structure used for their intervals.

Structural evidence validity. $V(E(x))$ requires a nonempty citation set plus table/key existence, concept and patient consistency, timestamp $\le t_0$, valid note offsets, and citation-hash integrity for every cited object; by convention $V(\varnothing) = 0$. Thus V is a pre-action provenance-and-integrity condition. Enforcing it gives zero acted-on failures of the listed checks by construction, whereas each observed post-t_0 citation under a removal policy witnesses $\widehat{R}_V > 0$. Semantic support is evaluated independently below.

Semantic evidence test. A blinded audit tests whether the complete valid citation set supports the decision, handles contradictions, and changes the adjudication when removed. Structural validity and semantic support answer different questions on the same evidence objects: the first establishes availability, identity, temporal validity, and integrity before routing; the second measures decision relevance on annotated held-out cases.

Shift protocol and monitor. Three prespecified triggers track FE above 1.0%, retrieval precision below 0.85, and coverage below 50%. A trigger pauses the affected stratum; a senior adjudicator reviews flagged cases and recent coordinator labels from the latest eight-week window; scenario-local recalibration follows; and routing resumes after FE and leakage checks pass. This protocol follows monitoring guidance for healthcare ML (Wiens et al., 2019; Chen et al., 2019; Lekadir et al., 2025; Fang et al., 2025).

Evaluation chronology. Stack development and the calibration partition precede policy locking. The primary α, all five sensitivity settings, and every comparator threshold are fixed before their single evaluation on the grouped holdout. Stress cohorts first receive the unchanged policies; scenario-local labels and recalibration enter only the separately reported recovery phase. The FHIR cohort likewise receives the original lock, with resource-native identifiers substituted inside the same evidence predicate. This chronology separates primary scoring, frozen transfer, monitor detection, and recovery.

4 Experimental Setup

Evaluation questions. The experiments ask three questions in sequence: does structural validity change FE and coverage at a fixed uncertainty event; do the routed citations carry decision-relevant content; and does the policy separation persist across operating points, record substrates, shifts, and integrity interventions?

Cohorts and split roles. The retrospective OMOP benchmark contains 84,291 patient–criterion pairs from 34 completed trials across 86 sites: 12 oncology, 8 cardiology, 6 neurology, 4 infectious-disease, and 4 other trials (Hripcsak et al., 2015). Each unit joins one atomic inclusion or exclusion criterion to a time-indexed chart at screening cutoff t_0. All criteria from the same patient–trial episode remain together in the 60/15/25 train/calibration/holdout partition, preventing episode overlap. The standard grouped holdout is distinct from three prespecified stress cohorts; the 12-site cohort is absent from stack development, policy selection, and standard scoring. A separate FHIR-native transfer cohort is absent from every OMOP role and receives the original policies.

Partition	Share/object	Permitted use	Reported role
Train	60% benchmark	Stack development	No final scoring
Calibration	15% benchmark	Lock α and all comparator thresholds	No result selection on test
Grouped holdout	25% benchmark	Final scoring only	Table 1, Figure 1
Stress cohorts	12 unseen sites; Q3 2025; OMOP v5.5	Frozen-policy scoring, then scenario-local recovery	Table 2, Figure 2

Frozen-stack closure and comparators. The matched policy family shares Meta-Llama-3.1-70B-Instruct with deterministic decoding, BGE-large-en-v1.5 retrieval over FAISS HNSW ($M = 32$, $\mathrm{efSearch} = 128$), bge-reranker-large, the same 80 KB context budget, and the same calibration data; only routing changes. Its five rows are singleton conformal plus V, the identical singleton event without V, V without singleton uncertainty, a calibration-tuned global score without V, and calibration-tuned directional FE thresholds without V. The first three form the component decomposition; the last two are score-routing references. LLM direct adjudication and rules/regex provide broader system references outside the matched family. Inference uses vLLM 0.5.5 with GPTQ 4-bit weights on 4×A100 80 GB GPUs; Appendix S1 gives the remaining configuration and latency profile.

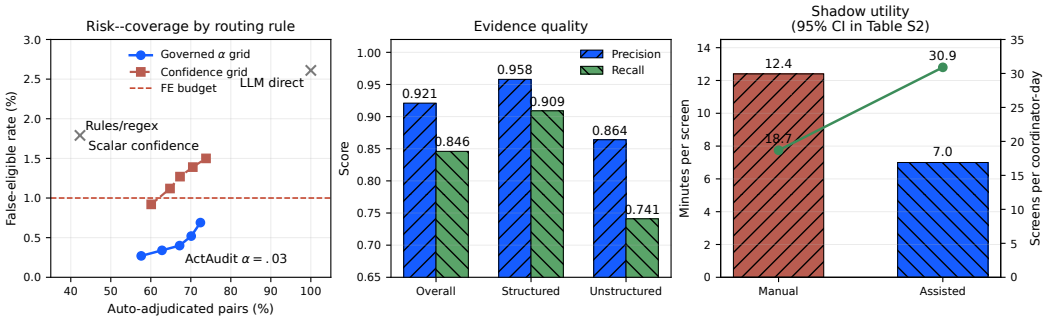

Figure 1: Three views of selective action. Left: conjunctive and scalar-confidence trajectories across prespecified operating points, compared at nearest coverage. Middle: precision/recall for structured and note-span evidence. Right: human-controlled shadow workflow endpoints, with distinct denominators in Table S2.

Labels, metrics, and uncertainty. Two certified clinical research coordinators independently label each pair, with senior resolution; the stratified double-annotation audit yields Cohen's $\kappa = 0.84$ (95% CI 0.82–0.86). FE denotes an automated inclusion-MET decision when gold is NOT-MET, or an exclusion-NOT-MET decision when gold is MET; false-ineligible (FI) is the directional converse. FE and FI divide by acted-on pairs, whereas patient-level resolution divides by patient–trial pairs. Inference and policy ties are deterministic; 1,000 trial- and site-stratified bootstrap replicates provide 95% intervals and paired policy contrasts. Evidence precision/recall uses 8,847 coordinator-annotated pairs, exact structured-record matching, and token-overlap matching for note spans after validity checks. The semantic study audits all 240 conjunctive-gate FE/FI errors plus a stratified sample of correct note-dependent actions with two qualified adjudicators and senior resolution. The non-interventional eight-week shadow study covers 1,126 screening events at 9 sites under mandatory human attestation, following silent-mode evaluation and clinical-AI reporting guidance (Kwong et al., 2022; McCradden et al., 2025; Liu et al., 2020). Appendix S2 gives the remaining protocols, stress definitions, and integrity-family sizes.

5 Results

5.1 Coverage-matched routing isolates the evidence effect

The closest coverage comparison is the most direct policy test: at 67.2% and 67.3% coverage, the conjunctive policy and scalar-grid point τ_3 have FE rates of 0.40% and 1.27%, with 0 and 11 post-t_0 citations, respectively (Table S5). Table 1 reports the prespecified primary scalar point τ_5 and decomposes the routing mechanism. Removing only V from the same singleton event adds 3.4 coverage points, 0.31 FE points, and 14 post-t_0 citations. Removing only singleton uncertainty adds 12.3 coverage points and 1.31 FE points. The primary scalar policy gains 6.6 coverage points but adds 1.10 FE points (95% CI 0.91–1.30) and 17 temporal violations; the directional FE-budgeted rule trades away 6.5 coverage points yet still exposes nine. The structural guarantee follows from enforcing V, while the paired FE changes show that this intervention also selects a clinically different error profile.

Figure 1 left tests whether the paired separation depends on one threshold choice. Across the five prespecified α values, conjunctive coverage spans 57.6–72.4% and FE spans 0.27–0.69%, with every acted-on citation passing V. Under nearest-coverage matching, scalar confidence has a minimum FE disadvantage of 0.65 points (95% CI 0.41–0.78); Table S5 reports the complete grid and matching rule.

5.2 Structural and semantic evidence tests agree on routing relevance

Evidence auditing on 8,847 coordinator-annotated pairs gives precision/recall 0.921/0.846 overall, 0.958/0.909 for structured records, and 0.864/0.741 for note spans. The independent semantic study asks whether structurally valid evidence affects the decision. Among 420 correct note-dependent

Table 1: Policy decomposition on 21,073 standard grouped-holdout pairs. The scalar-confidence row is the prespecified primary point τ_5; the coverage-matched τ_3 comparison appears in Table S5. The first five rows share all prediction tuples; the last two are broad system references. "Yes" means V is enforced over source, nonemptiness, concept, patient, time, offset, and integrity. FE/FI intervals and paired differences use 1,000 grouped bootstrap replicates.

System / policy	Auto n (%)	FE n (%; 95% CI)	FI n (%; 95% CI)	V enforced; post-t_0/cited
ACTAUDIT conjunctive gate	14,162 (67.2)	57 (0.40; 0.30–0.52)	183 (1.29; 1.12–1.48)	Yes; 0/129,844
Scalar confidence (τ_5)	15,552 (73.8)	233 (1.50; 1.31–1.70)	462 (2.97; 2.70–3.24)	No; 17/126,321
Conformal-only	14,874 (70.6)	106 (0.71; 0.58–0.86)	263 (1.77; 1.56–1.98)	No; 14/132,906
Validity-only	16,751 (79.5)	286 (1.71; 1.51–1.91)	529 (3.16; 2.89–3.43)	Yes; 0/153,918
FE-budgeted dual threshold	12,792 (60.7)	102 (0.80; 0.65–0.96)	271 (2.12; 1.87–2.37)	No; 9/104,876
LLM direct adjudication	21,073 (100.0)	550 (2.61; 2.40–2.83)	1,018 (4.83; 4.55–5.12)	No; 401/134,775
Rules/regex phenotype	8,914 (42.3)	160 (1.79; 1.53–2.08)	701 (7.86; 7.30–8.44)	–

Policy minus conjunctive gate	ΔFE points (95% CI)	Δcoverage points (95% CI)
Scalar confidence	+1.10 (+0.91–+1.30)	+6.6 (+5.9–+7.3)
Conformal-only	+0.31 (+0.16–+0.46)	+3.4 (+2.7–+4.1)
Validity-only	+1.31 (+1.13–+1.49)	+12.3 (+11.5–+13.1)
FE-budgeted dual threshold	+0.40 (+0.25–+0.56)	-6.5 (-7.2 to -5.8)

actions, 356 (84.8%; 95% CI 81.1–88.0) have citation sets judged sufficient after resolution, and removing the cited set changes the label or forces deferral in 371/420 cases (88.3%). Table S6 further assigns every FE/FI error to a resolved mechanism: evidence unavailability and retrieval misses account for 26/57 FE and 83/183 FI errors, whereas adjudicator reasoning accounts for 5/57 FE and 21/183 FI errors. The routing intervention and semantic audit therefore converge on evidence access as a distinct error mechanism.

5.3 DISTRIBUTION SHIFT CHANGES THE EMPIRICAL ACTION REGION

The first panel of Table 2 applies both frozen policies to 2,921 pairs from 12 unseen sites. The conjunctive gate reaches 1.31% site-weighted FE with zero leakage, outside $\widehat{\mathcal{A}}_{\epsilon,\gamma}$, while scalar confidence reaches 2.36% FE with 8 post-cutoff citations. Their paired FE difference is +1.05 points (95% CI +0.61–+1.51). The remaining panels report scenario-local post-recalibration measurements: FE is below 0.6% and coverage above 61% in all three affected strata. Vocabulary migration is mechanistically different: retrieval precision falls to 0.812 and triggers the evidence branch while FE remains below its ceiling, showing that FE alone would miss a material change in the routed evidence.

Pair-level replay over the shadow window produces 17 alarms (0.38 per 1,000 evaluated patient–criterion pairs; 0.94 per site-month), of which 3 (17.6%) are false alarms under the prespecified no-breach/no-action-change definition. Resolution uses a median 74 fresh labels (IQR 58–96) and 3.6 senior-review hours (p90 6.8); four-week post-reenable FE/coverage are 0.48%/63.6%. These measurements quantify both detection frequency and the human work attached to a pause.

Two integrity suites reach the evidence-eligibility boundary from an adversarial direction. In 2,640 prespecified attacks, 2,421 are blocked before inference, 218 are deferred, and one note-injection attack produces unsafe auto-action; delimiter normalization then yields zero unsafe actions across 1,140 replays. In a separate set of 1,000 held-out mutations, the full gate permits 2 unsafe actions (0.20%), compared with 102 (10.2%) across family-specific weakened-check controls. Benign false

Table 2: Selective action under stress. The first panel compares frozen policies on 12 unseen sites using equal-site-weighted rates and pooled raw counts; weighted percentages therefore need not equal pooled count ratios. Subsequent panels report same-stratum post-recalibration measurements, separately from the untouched primary and FHIR evaluations.

Frozen policy, 12 unseen sites	Auto n (weighted %)	FE n (weighted %; 95% CI)	Leakage / cited
Conjunctive gate	1,878 (64.3)	25 (1.31; 0.89–1.88)	0/17,204
Scalar confidence	2,057 (70.4)	49 (2.36; 1.73–3.05)	8/16,991

Recovery scenario	n pairs	FE pre → post (%)	Coverage pre → post (%)	Pause-to-reenable
Grouped-holdout reference (no recovery)	21,073	0.40 → 0.40	67.2 → 67.2	–
Site-held-out (12 unseen sites)	2,921	1.31 → 0.58	64.3 → 61.7	12.4 h
Temporal shift (Q3 2025 notes)	3,444	1.07 → 0.49	69.8 → 66.9	10.9 h
Vocabulary shift (OMOP v5.4 → v5.5)	1,876	0.94 → 0.44	65.5 → 64.1	11.5 h

Scenario	Trigger branch	Retrieval precision pre → post
Site-held-out	FE + retrieval	0.827 → 0.902
Temporal	FE	0.861 → 0.913
Vocabulary	Retrieval	0.812 → 0.906

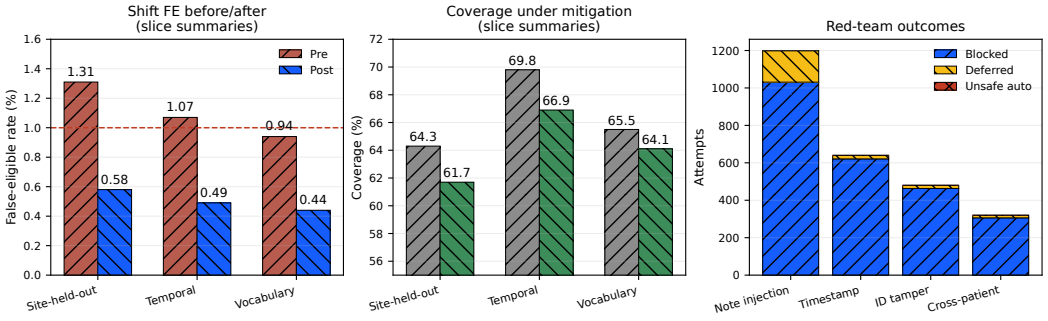

Figure 2: Frozen shift and integrity stress. Left and middle: same-stratum post-recalibration measurements lower FE from 1.31/1.07/0.94% to 0.58/0.49/0.44% while retaining more than 61% coverage. Right: blocked, deferred, and unsafe outcomes for four prespecified attacks on structural evidence validity.

deferral is 43/6,220 (0.69%); median/p95 validation overhead is 7.8/18.6 ms. Tables S8 and S3 trace each threat to its protected evidence property.

Patient-level workflow utility. On 3,984 held-out patient–trial pairs, ACTAUDIT fully resolves 1,671 (41.9%; 95% CI 40.4–43.5); 2,313 (58.1%) retain at least one deferred criterion, with median 2 unresolved criteria (IQR 1–4; p90 5). In the human-controlled shadow study, the observed median coordinator time is 7.0 minutes per screen, 43.5% lower (95% CI 40.2–46.8) than the 12.4-minute immediately preceding comparison window; separately estimated case-mix-weighted throughput is 30.9 versus 18.7 screens per coordinator-day. Override, evidence-edit, and independent discordance rates are 11.4%, 15.8%, and 34/4,149 (0.82%), respectively. This observational workflow association complements the criterion-level routing results while retaining mandatory coordinator attestation.

Substrate replication. An untouched FHIR-native oncology eligibility service spanning 6,240 pairs, 9 trials, and 4 institutions tests the policy effect under different record identifiers. With the original lock, the conjunctive gate achieves 62.9% coverage at 0.61% FE, compared with 71.7%

coverage and 1.68% FE for scalar confidence. The nearer-coverage conformal-only policy reaches 66.4% coverage at 0.84% FE and admits nine post-cutoff citations. The five-policy coverage ordering and joint FE/temporal-validity separation persist after substituting FHIR-native resource identifiers; Table S9 reports raw counts, FI, citations, evidence quality, and grouped intervals.

6 DISCUSSION

The paired interventions establish that uncertainty and structural evidence validity select different examples. In the binary setting, conformal-only is already a score threshold derived from calibration. Adding V to that identical event removes 3.4 coverage points, 0.31 FE points, and all 14 observed temporal violations; removing singleton uncertainty from V instead adds 12.3 coverage points and 1.31 FE points. The result is therefore an interaction between two routing coordinates on fixed predictions, not a gain from changing the retriever or adjudicator.

This distinction sharpens the contribution relative to conformal RAG and clinical evidence gates. Conformal selection controls an uncertainty or learned-quality event; evidence grading assesses whether content warrants a claim. ACTAUDIT makes structural evidence history directly observable before routing and then measures semantic support on the resulting actions. That decomposition explains why the matched score at 67.3% coverage still reaches 1.27% FE with temporal violations, while the conjunction at 67.2% reaches 0.40% FE without them.

The semantic, transfer, and integrity studies test the same mechanism from complementary directions. Citation sets are judged sufficient for 84.8% of audited correct note-dependent actions, and removing them changes the label or forces deferral in 88.3%. FHIR transfer preserves the policy ordering and FE separation after replacing OMOP keys with resource-native identifiers. Held-out integrity mutations reduce unsafe action from 10.2% under family-specific weakened checks to 0.20% under the full gate, while the 0.69% benign false-deferral rate and millisecond-scale validation overhead quantify the cost of enforcing the predicate.

The evidence supports criterion-level routing in episode-grouped retrospective cohorts and frozen FHIR transfer. Patient–trial episode is the operational split unit, and trial- and site-stratified intervals target the multi-site design; repeated individuals in distinct episodes are therefore represented at the episode level. V certifies structural provenance and integrity, with semantic sufficiency measured as a second layer on note-dependent actions. Post-recalibration panels characterize recovery in the affected strata, and the shadow time comparison is an observational, case-mix-weighted association under human attestation. Accordingly, the strongest inference comes from the paired grouped-holdout interventions and their untouched FHIR replication.

7 CONCLUSION

ACTAUDIT separates label generation from evidence-conditioned action on a shared clinical RAG stack. At nearly equal coverage on 21,073 held-out pairs, joining singleton uncertainty with structural evidence validity reduces FE from 1.27% to 0.40% and post-cutoff citations from 11 to zero. Exact component removals, five prespecified operating points, and frozen FHIR transfer at 0.61% versus 1.68% FE reproduce the separation from complementary angles. Structural provenance and integrity therefore form an actionable routing coordinate, with semantic support independently confirmed by blinded evidence-removal tests; measuring both changes which clinical trial screening predictions advance without coordinator review.

ETHICS STATEMENT

The retrospective benchmark used site-specific ethics approvals and/or exemptions with waiver of consent as locally required. The 9-site shadow deployment was non-interventional, and no screening decision could be submitted without human attestation. Controls address automation bias, documentation-driven inequity, PHI exposure, and misuse through required coordinator review, logged overrides and evidence edits, per-citation validation, and stratum-level pause policies (Wiens et al., 2019; McCradden et al., 2020; Lekadir et al., 2025).

Reproducibility Statement

Sections 3–5 specify the evaluation target, routing rule, split roles, matched controls, estimands, uncertainty procedure, and decision thresholds. Appendix S1 records the complete model, retrieval, serving, hardware, and runtime configuration; Appendix S2 reports sensitivity grids, annotation protocols, raw counts, monitor burden, transfer results, and adversarial controls needed to reproduce every reported comparison.

AI Use Statement

The manuscript-preparation record documents that generative AI tools assisted with iterative drafting, review, and submission preparation, including organization and editing of the text, title, abstract, and presentation.

References

Nour Alkhoury, Maqsood Shaik, Ricardo Wurmus, and Altuna Akalin. Enhancing biomarker based oncology trial matching using large language models. *npj Digital Medicine*, 8(1):250, 2025. doi: 10.1038/s41746-025-01673-4.

Rami Aly, Zhiqiang Tang, Samson Tan, and George Karypis. Learning to generate answers with citations via factual consistency models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11876–11896. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.641. URL https://aclanthology.org/2024.acl-long.641/.

A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Chia-Yuan Chang, Jiayi Yuan, Sirui Ding, Qiaoyu Tan, Kai Zhang, Xiaoqian Jiang, Xia Hu, and Na Zou. Towards fair patient-trial matching via patient-criterion level fairness constraint. In *AMIA Annual Symposium Proceedings*, pp. 884–893, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10785912/.

P.-H. C. Chen, Y. Liu, and L. Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18:410–414, 2019. doi: 10.1038/s41563-019-0345-0.

Zhongnan Fang, Andrew Johnston, Lina Y. Cheuy, Hye Sun Na, Magdalini Paschali, Camila Gonzalez, Bonnie A. Armstrong, Arogya Koirala, Derrick Laurel, Andrew Walker Campion, Michael Iv, Akshay S. Chaudhari, and David B. Larson. Automated real-time assessment of intracranial hemorrhage detection AI using an ensembled monitoring model (EMM). *npj Digital Medicine*, 8 (1):608, 2025. doi: 10.1038/s41746-025-02007-0.

Naihe Feng, Yi Sui, Shiyi Hou, Jesse C. Cresswell, and Ga Wu. Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2832–2836. ACM, 2025. doi: 10.1145/3726302.3730244.

D. B. Fogel. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*, 11:156–164, 2018. doi: 10.1016/j.conctc.2018.08.001.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and

revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

T. Gao, H. Yen, J. Yu, and D. 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, 2023b. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

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

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

K. A. Getz and R. A. Campo. Trial watch: Trends in clinical trial design complexity. *Nature Reviews Drug Discovery*, 16(5):307, 2017. doi: 10.1038/nrd.2017.65.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Advances in Neural Information Processing Systems*, volume 37, pp. 73884–73919, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/870ccde24673d3970a680bb48496ed63-Abstract-Conference.html.

Shashi Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N. Kothari, Regina Schwind, Therica M. Miller, Sorena Nadaf-Rahrov, Yanshan Wang, and Hrituraj Singh. PRISM: Patient records interpretation for semantic clinical trial matching system using large language models. *npj Digital Medicine*, 7(1):305, 2024. doi: 10.1038/s41746-024-01274-7.

G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20:117–121, 2013. doi: 10.1136/amiajnl-2012-001145.

George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. In *MEDINFO 2015: eHealth-Enabled Health*, volume 216 of *Studies in Health Technology and Informatics*, pp. 574–578. IOS Press, 2015. doi: 10.3233/978-1-61499-564-7-574. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4815923/.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14095–14113, 2024. doi: 10.18653/v1/2024.findings-acl.838. URL https://aclanthology.org/2024.findings-acl.838/.

Qiao Jin, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *Nature Communications*, 15(1):9074, 2024. doi: 10.1038/s41467-024-53081-z.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000. PMLR, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W. Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071, 2017. doi: 10.1093/jamia/ocx019.

Jethro C. C. Kwong, Lauren Erdman, Adree Khondker, Marta Skreta, Anna Goldenberg, Melissa D. McCradden, Armando J. Lorenzo, and Mandy Rickard. The silent trial—the bridge between bench-to-bedside clinical AI applications. *Frontiers in Digital Health*, 4:929508, 2022. doi: 10.3389/fdgth.2022.929508.

Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, Ben Glocker, Celia Cintas, Curtis P. Langlotz, Eva Weicken, Folkert W. Asselbergs, Fred Prior, Gary S. Collins, Georgios Kaissis, Gianna Tsakou, Irène Buvat, Jayashree Kalpathy-Cramer, John Mongan, Julia A. Schnabel, Kaisar Kushibar, Katrine Riklund, Kostas Marias, Lameck M. Amugongo, Lauren A. Fromont, Lena Maier-Hein, Leonor Cerdá-Alberich, Luis Martí-Bonmatí, M. Jorge Cardoso, Maciej Bobowicz, Mahsa Shabani, Manolis Tsiknakis, Maria A. Zuluaga, Marie-Christine Fritzsche, Marina Camacho, Marius George Linguraru, Markus Wenzel, Marleen De Bruijne, Martin G. Tolsgaard, Melanie Goisauf, Mónica Cano Abadía, Nikolaos Papanikolaou, Noussair Lazrak, Oriol Pujol, Richard Osuala, Sandy Napel, Sara Colantonio, Smriti Joshi, Stefan Klein, Susanna Aussó, Wendy A. Rogers, Zohaib Salahuddin, Martijn P. A. Starmans, and FUTURE-AI Consortium. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, 2025. doi: 10.1136/bmj-2024-081554.

Daniel León Tramontini, Shrestha Ghosh, and Carsten Eickhoff. Investigating RAG-based approaches in clinical trial and patient matching. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 76–87. PMLR, 2026. URL https://proceedings.mlr.press/v297/leon-tramontini26a.html.

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. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. 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. 3799–3821. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.210. URL https://aclanthology.org/2024.naacl-long.210/.

Che Lin, Jia-Yi Lin, and Yao-San Lin. Evidence-graded decision authorization for safe clinical AI: A constrained reasoning framework. *medRxiv*, 2026. doi: 10.64898/2026.05.19.26353565. URL https://www.medrxiv.org/content/10.64898/2026.05.19.26353565v1.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, and SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020. doi: 10.1016/S2589-7500(20)30218-1.

M. D. McCradden, E. A. Stephenson, and J. A. Anderson. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nature Medicine*, 26:1325–1326, 2020. doi: 10.1038/s41591-020-1035-9.

Melissa D. McCradden, Alex John London, Judy Wawira Gichoya, Mark Sendak, Lauren Erdman, Ian Stedman, Lauren Oakden-Rayner, Ismail Akrout, James A. Anderson, Lesley-Anne Farmer, Robert Greer, Anna Goldenberg, Yvonne Ho, Shalmali Joshi, Jennie Louise, Muhammad Mamdani, Mjaye L. Mazwi, Abdullahi Mohamud, Lyle J. Palmer, Antonios Peperidis, Stephen R. Pfohl, Mandy Rickard, Carolyn Semmler, Karandeep Singh, Devin Singh, Seyi Soremekun, Lana Tikhomirov, Anton H. van der Vegt, Karin Verspoor, and Xiaoxuan Liu. CANAIRI: The collaboration for translational artificial intelligence trials in healthcare. *Nature Medicine*, 31(1):9–11, 2025. doi: 10.1038/s41591-024-03364-1.

R. Miotto and C. Weng. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *Journal of the American Medical Informatics Association*, 22: e141–e150, 2015. doi: 10.1093/jamia/ocu050.

Yizhao Ni, Stephanie Kennebeck, Judith W. Dexheimer, Constance M. McAneney, Huaxiu Tang, Todd Lingren, Qi Li, Haijun Zhai, and Imre Solti. Automated clinical trial eligibility prescreening: Increasing the efficiency of patient identification for clinical trials in the emergency department. *Journal of the American Medical Informatics Association*, 22(1):166–178, 2015. doi: 10.1136/amiajnl-2014-002887.

M. Nievas, A. Basu, Y. Wang, and H. Singh. Distilling large language models for matching patients to clinical trials. *Journal of the American Medical Informatics Association*, 31:1953–1963, 2024. doi: 10.1093/jamia/ocae073.

H. Qiu, E. Dobriban, and E. Tchetgen Tchetgen. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society: Series B*, 85:1680–1705, 2023. doi: 10.1093/jrsssb/qkad069.

Y. Romano, M. Sesia, and E. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html.

Jessica Ross, Samson W. Tu, Simona Carini, and Ida Sim. Analysis of eligibility criteria complexity in clinical trials. *Summit on Translational Bioinformatics*, 2010:46–50, 2010. URL https://pubmed.ncbi.nlm.nih.gov/21347148/.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J. Embi, Noémie Elhadad, Stephen B. Johnson, and Albert M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2): 221–230, 2014. doi: 10.1136/amiajnl-2013-001935.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, 2019. doi: 10.1038/s41591-019-0548-6.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium Proceedings*, pp. 1324–1333, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10785941/.

SUPPLEMENTARY MATERIAL

S1 IMPLEMENTATION AND PROTOCOL DETAILS

Model and retrieval stack. Criterion decomposition and adjudication use Meta-Llama-3.1-70B-Instruct with deterministic decoding. Adjudication is served in GPTQ 4-bit format through vLLM 0.5.5. Retrieval uses BGE-large-en-v1.5 embeddings with a FAISS HNSW index ($M = 32$, $\mathrm{efSearch} = 128$), followed by bge-reranker-large. Every policy in the matched comparison receives the same retrieved candidates, reranked context, 80 KB input budget, normalized MET/NOT-MET scores, and deterministic output. Split conformal ranks these scores against the shared calibration partition; probability calibration is not assumed.

Compute and runtime. Experiments run on Ubuntu 22.04.4 LTS, CUDA 12.4, cuDNN 9.1, PyTorch 2.2.2, FAISS 1.8.0, and Python 3.11.8. The production-like cluster provides 4×NVIDIA A100 80 GB SXM4 GPUs, 64-core AMD EPYC 7V13 CPUs, and 512 GB RAM. Median end-to-end latency is 3.9 s per patient–criterion pair and p95 latency is 10.8 s. Meta-Llama-3.1 is used under the Llama 3.1 Community License; FlagEmbedding and FAISS use MIT licenses; vLLM uses Apache 2.0; and OMOP vocabularies follow OHDSI/Athena source terms.

Annotation and evidence rubric. Certified coordinators first identify the criterion polarity and time window, then assign MET/NOT-MET, mark structured records or minimal note spans supporting the decision, and record any contradiction. A second coordinator works independently. A senior adjudicator resolves label, evidence, and temporal disagreements. Structured citations match table, key, concept, and value; note citations match source note, offsets, timestamp, and overlapping support tokens after validity checks. The sufficiency extension asks whether the complete cited set justifies the label, whether cross-source contradictions are resolved correctly, and whether evidence removal changes the label or forces deferral.

Shadow workflow. The 9-site shadow deployment is non-interventional. Coordinators inspect the proposed label, open the linked evidence chain, edit any label or citation when needed, and attest that the screen reflects their judgment before submission. No routed output can be submitted automatically. The observational comparison uses the immediately preceding eight-week period and case-mix weighting by therapeutic area and criterion count. Review time per screen and screens per coordinator-day are estimated separately under this weighting.

S2 COHORT HETEROGENEITY AND WORKFLOW ENDPOINTS

Figure S1 and Table S1 summarize heterogeneity within the standard grouped benchmark holdout. They aggregate sites and trials represented in that holdout; they are distinct from the additional 12-site frozen-policy stress cohort in Table 2. Thus the 0.89% site-level 90th percentile below is a within-benchmark distribution statistic, whereas the 1.31% pre-repair value in Table 2 is an equal-site-weighted estimate on unseen sites.

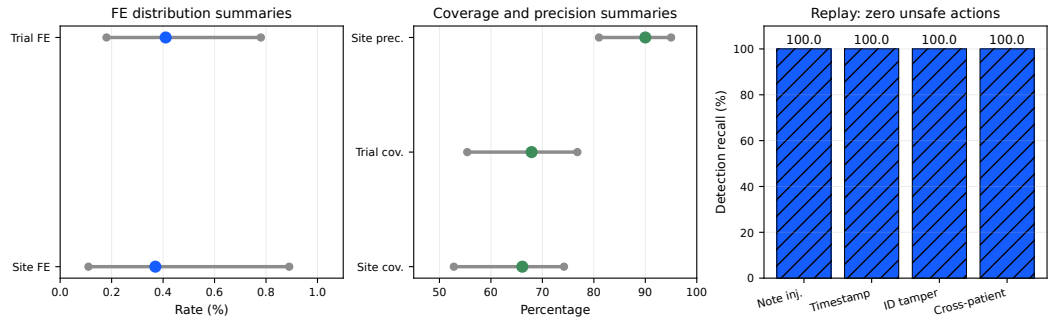

Figure S1: Cohort and replay summaries. Left: site- and trial-level FE percentiles within the standard grouped holdout. Middle: coverage and site-level evidence-precision percentiles on the same evaluation object. Right: delimiter-normalized replay detects all 1,140 mutations with zero unsafe action; independent held-out mutations appear in Table S8.

Table S1: Within-benchmark site- and trial-level distributions on the standard grouped holdout. These aggregations exclude the separate 12-site stress cohort in Table 2.

Metric	10th percentile	Median	90th percentile
Site-level FE rate (%)	0.11	0.37	0.89
Site-level coverage (%)	52.8	66.1	74.2
Site-level evidence precision	0.81	0.90	0.95
Trial-level FE rate (%)	0.18	0.41	0.78
Trial-level coverage (%)	55.4	67.9	76.8

The within-benchmark 90th-percentile FE rate remains below 1% at both site and trial level, while the 10th-percentile coverage remains above 50%. These distribution summaries support the aggregate primary point without conflating represented benchmark sites with the unseen-site stress cohort.

Table S2: Patient-level endpoints and observed operational associations from the human-controlled 9-site shadow deployment. Throughput and time per screen are separate case-mix-weighted estimands.

Metric	Value	95% CI / detail
Fully auto-resolved patient–trial pairs	1,671/3,984 (41.9%)	[40.4, 43.5]
Pairs with at least one deferred criterion	2,313/3,984 (58.1%)	[56.5, 59.6]
Median deferred criteria (deferred pairs only)	2	IQR 1–4; p90 = 5
Manual baseline review time per screen	12.4 min	[11.8, 13.1]
Assisted shadow review time per screen	7.0 min	[6.6, 7.4]
Relative time reduction	43.5%	[40.2, 46.8]
Coordinator override rate	11.4%	[10.0, 12.9]
Evidence-chain edit rate	15.8%	[14.1, 17.6]
Independent audit discordance	0.82%	34/4,149 reviewed criteria

The shadow endpoints retain human attestation as the unit of workflow use: the model fully resolves 41.9% of patient–trial pairs, and the assisted window has a 43.5% lower separately estimated median time per screen than the preceding observational comparison window.

Table S3: Red-team integrity evaluation over 2,640 prespecified attacks.

Attack family	Attempts	Blocked	Deferred	Unsafe auto	Detection recall
Instruction injection in note text	1,200	1,031	168	1	1,199/1,200 (99.9)
Timestamp backfill manipulation	640	621	19	0	640/640 (100.0)
Evidence ID swap / tamper	480	463	17	0	480/480 (100.0)
Cross-patient record stitching	320	306	14	0	320/320 (100.0)
Total	2,640	2,421	218	1	2,639/2,640 (>99.9)

The original suite leaves one unsafe note-injection action among 2,640 attempts; delimiter normalization targets that parser failure, and Table S4 reports the prespecified replay by attack family.

Table S4: Delimiter-normalized replay by attack family.

Attack family (replay)	Attempts	Unsafe auto	Detection recall
Instruction injection in note text	600	0	600/600 (100.0)
Timestamp backfill manipulation	220	0	220/220 (100.0)
Evidence ID swap / tamper	180	0	180/180 (100.0)
Cross-patient record stitching	140	0	140/140 (100.0)
Total	1,140	0	1,140/1,140 (100.0)

All 1,140 replayed mutations are detected without an unsafe action. The independent held-out suite in Appendix S6 prevents this targeted replay from serving as the only integrity result.

S3 Risk–Coverage Behavior Across Prespecified Operating Points

Table S5 scores five α values and a scalar-score grid, all fixed before evaluation on the same 21,073-pair grouped holdout. The primary design point is $\alpha = 0.03$; the neighboring settings characterize local risk–coverage behavior and are not candidates for revising that lock. The conjunctive trajectory remains inside the FE $\leq 1.0\%$ and coverage $\geq 50\%$ target region at every setting. For a comparable-coverage contrast, each conjunctive point is paired with the scalar-confidence point minimizing absolute coverage difference, and the minimum FE gap is taken over these five descriptive pairs.

Table S5: Prespecified action grid. Coverage and FE intervals use 1,000 trial- and site-stratified bootstrap replicates. Conjunctive rows enforce the complete structural predicate V; scalar-confidence rows do not, and their post-t_0 counts are exact over policy-specific cited objects.

Policy	Setting	Auto n (%; 95% CI)	FE n (%; 95% CI)	Leakage / cited
Conjunctive	$\alpha = 0.01$	12,138 (57.6; 56.8–58.4)	33 (0.27; 0.19–0.38)	0/111,216
Conjunctive	$\alpha = 0.02$	13,235 (62.8; 62.0–63.6)	45 (0.34; 0.25–0.46)	0/121,038
Conjunctive	$\alpha = 0.03$	14,162 (67.2; 66.4–68.0)	57 (0.40; 0.30–0.52)	0/129,844
Conjunctive	$\alpha = 0.04$	14,773 (70.1; 69.3–70.9)	77 (0.52; 0.41–0.66)	0/135,289
Conjunctive	$\alpha = 0.05$	15,257 (72.4; 71.6–73.2)	105 (0.69; 0.56–0.83)	0/139,904
Scalar confidence	τ_1	12,667 (60.1; 59.3–60.9)	117 (0.92; 0.76–1.10)	5/103,874
Scalar confidence	τ_2	13,657 (64.8; 64.0–65.6)	153 (1.12; 0.95–1.31)	9/111,628
Scalar confidence	τ_3	14,183 (67.3; 66.5–68.1)	180 (1.27; 1.09–1.47)	11/115,436
Scalar confidence	τ_4	14,857 (70.5; 69.7–71.3)	207 (1.39; 1.21–1.59)	14/120,755
Scalar confidence	τ_5	15,552 (73.8; 73.1–74.5)	233 (1.50; 1.31–1.70)	17/126,321

At the nearly identical 67.2% and 67.3% coverage points, the conjunctive and confidence policies have FE rates of 0.40% and 1.27%, respectively. This pair makes the local separation visible without relying on the more distant primary confidence operating point.

S4 Semantic Support and Residual Attribution

Two independent qualified adjudicators evaluate whether citation sets carry sufficient semantic support and resolve cross-source contradictions without seeing routing confidence. Senior resolution produces the final labels. The counterfactual removal test reruns the adjudicator after deleting the cited set while leaving the remaining chart context unchanged.

Table S6: Semantic support audit and exhaustive attribution of all conjunctive-gate errors. Percentages in the lower panel use 57 FE and 183 FI errors as denominators.

Note-evidence endpoint	Result	95% CI / denominator
Correct note-dependent auto-actions audited	420	Stratified held-out sample
Citation set sufficient after resolution	356 (84.8%)	81.1–88.0
Cross-source contradiction handled correctly	78/96 (81.3%)	96 contradiction-bearing actions
Evidence removal changes label or defers	371/420 (88.3%)	85.0–91.2
Inter-adjudicator agreement	$\kappa = 0.82$	0.78–0.86

Resolved primary attribution	FE n (%)	FI n (%)
Evidence unavailable in chart	15 (26.3)	49 (26.8)
Retrieval miss	11 (19.3)	34 (18.6)
Temporal-window reasoning	10 (17.5)	23 (12.6)
Cross-source contradiction	7 (12.3)	19 (10.4)
Ambiguous clinical language	6 (10.5)	25 (13.7)
Adjudicator reasoning error	5 (8.8)	21 (11.5)
Outside-system missingness	3 (5.3)	12 (6.6)
Total	57 (100.0)	183 (100.0)

Evidence removal changes the adjudication or forces deferral in 371/420 cases, while the attribution panel assigns 109/240 governed errors to unavailable evidence or retrieval misses. These outcomes distinguish decision relevance from the deterministic structural checks in V.

S5 Shift Triggers, Review Burden, and Post-Recalibration Measurements

A false alarm is a monitor pause for which senior review finds no threshold breach, no invalid acted-on citation, and no action-changing relabel. Pair-level alarm rates divide by all continuously evaluated patient–criterion pairs derived from the 1,126 shadow screening events and by 18.1 site-months. Fresh-label counts include only labels acquired after alarm; review hours include senior chart review, resolution, and threshold approval. Post-reenable rates are same-stratum measurements after local recalibration and are not used to select or revise the untouched primary or FHIR results.

Table S7: Eligibility-revocation triggers, operational burden, and four-week stability after re-enablement.

Endpoint	Result	Detail
Retrieval precision, unseen-site pre → post	0.827 → 0.902	FE + retrieval trigger
Retrieval precision, temporal pre → post	0.861 → 0.913	FE trigger
Retrieval precision, vocabulary pre → post	0.812 → 0.906	Retrieval trigger
Routine alarms	17	0.38/1,000 pairs; 0.94/site-month
False alarms	3 (17.6%)	Prespecified no-breach/no-action-change definition
Fresh labels per alarm	Median 74	IQR 58–96
Senior review per alarm	Median 3.6 h	p90 6.8 h
Four-week post-reenable FE / coverage	0.48% / 63.6%	All re-enabled strata

The vocabulary scenario activates on retrieval precision even though pre-recalibration FE is below 1%, demonstrating that the evidence trigger detects a change not represented by the directional-error coordinate alone.

S6 Integrity Interventions and Weakened-Check Controls

The security evaluation treats clinical evidence objects and their authorization metadata as protected objects. Attackers may control note text or submitted identifiers but cannot alter the trusted patient-index service, server-side clock, keyed citation hashes, or access-control layer. Each family-specific comparator removes the targeted evidence check while retaining the same retriever and adjudicator.

Table S8: Threat model, evidence-eligibility boundary, held-out mutations, and family-specific weakened-check controls. Full-gate outcomes are blocked/deferred/unsafe; weakened-check cells report unsafe auto-actions.

Protected object	Attacker capability	Trust boundary and security property
Note span and parser boundary	Embed instructions and mutate delimiters	Note text is untrusted; preserve source identity, span validity, and control/data separation
Evidence timestamp	Backfill or alter apparent event time	Server clock and signed ingestion time are trusted; enforce evidence availability by t_0
Evidence identifier and hash	Swap keys or tamper with payload	Patient index and keyed hash are trusted; bind each citation identifier to immutable content
Patient ownership boundary	Stitch valid records across patients	Access-control patient index is trusted; enforce subject consistency before action

Family / strongest comparator	Adaptive or held-out setting	Observed failure, mitigation, and evidence trace
Note injection; disable delimiter and offset validation	Original adaptive injection suite plus held-out delimiter mutations	One original unsafe parse; delimiter normalization; Tables S3–S4
Timestamp backfill; disable temporal consistency	Prespecified suite plus held-out timestamp mutations	No original unsafe action; signed-time validation retained; Figure 2
Evidence ID tamper; disable keyed-hash validation	Prespecified suite plus held-out key/payload mutations	One held-out unsafe path; keyed binding and deferral; numeric panel below
Cross-patient stitching; disable patient-key consistency	Prespecified suite plus held-out record combinations	No original unsafe action; patient-key binding retained; Table S3

Held-out family	Mutations	Full gate: blocked / deferred / unsafe (unsafe %)	Weak gate: unsafe n (%)
Note injection	360	326/33/1 (0.28)	31 (8.61)
Timestamp backfill	240	227/13/0 (0.00)	27 (11.25)
Evidence ID tamper	220	207/12/1 (0.45)	25 (11.36)
Cross-patient stitching	180	168/12/0 (0.00)	19 (10.56)
Total	1,000	928/70/2 (0.20%)	102 (10.2)

Benign utility / overhead endpoint	Value	Denominator
Benign false deferral from integrity validation	43 (0.69%)	6,220 benign acted-on candidates
Validation latency	7.8 ms median	18.6 ms p95

Family-specific weakened checks produce 102 unsafe actions across 1,000 held-out mutations, compared with two under the full structural gate. The 0.69% benign false-deferral rate and millisecond-scale validation overhead quantify the utility cost of those checks on the tested candidates.

S7 Substrate Replication Beyond OMOP

The replication cohort is an untouched FHIR-native oncology eligibility service at four institutions. Its nine trials and 6,240 patient–criterion pairs are absent from OMOP stack development, calibration, and standard scoring. Each policy retains the original calibration lock, FE definition, evidence-validity

predicate, deterministic decoder, and bootstrap protocol. FHIR resources replace OMOP table/key references inside $V(E)$ while preserving source identity, timestamp, span, patient binding, and integrity requirements.

Table S9: Frozen-policy substrate replication on an untouched FHIR-native oncology cohort. Intervals are grouped 95% bootstrap intervals. "Yes" means the complete FHIR-native structural predicate V is enforced; otherwise the displayed post-cutoff citations witness a validity failure.

Policy	Auto n (%)	FE n (%; 95% CI)	FI n (%; 95% CI)	V enforced; post-t_0/cited	Evidence P/R
Conjunctive gate	3,925 (62.9)	24 (0.61; 0.39–0.91)	71 (1.81; 1.43–2.25)	Yes; 0/35,040	0.903/0.821
Scalar confidence	4,474 (71.7)	75 (1.68; 1.32–2.10)	139 (3.11; 2.62–3.64)	No; 11/36,620	0.889/0.835
Conformal-only	4,143 (66.4)	35 (0.84; 0.59–1.17)	89 (2.15; 1.73–2.62)	No; 9/35,970	0.895/0.827
Validity-only	4,724 (75.7)	91 (1.93; 1.55–2.35)	168 (3.56; 3.05–4.09)	Yes; 0/42,380	0.900/0.844
FE-budgeted dual threshold	3,451 (55.3)	27 (0.78; 0.51–1.13)	78 (2.26; 1.79–2.78)	No; 6/28,905	0.907/0.797

The substrate change preserves the policy ordering and the decisive coordinates: the conjunctive gate remains below 1% FE with no structural-validity failure, while every policy that omits V exposes post-cutoff citations. At similar coverage, FHIR-native conformal-only reaches 66.4% coverage and 0.84% FE, compared with 62.9% and 0.61% for the conjunction.