ADMITSCOPE: COMPILING EVIDENCE-GROUNDED LEADERBOARDS INTO AUDIT-QUALIFIED CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence-grounded leaderboards guide scientific comparison, but their score margins are trustworthy only when abstention policies, evaluated populations, and cited documents support the same conclusion. ADMITSCOPE makes that validity condition executable. Given a proposed row comparison, it compiles task scope, evidence access, coverage policy, metric, and uncertainty into a claim; activates obligations for paired margins, population comparability, document grounding, held-out stability, and independent review; and issues a direct, transferred, or surrogate-only license with an inspectable evidence trace. We instantiate ADMITSCOPE on 10,261 eligibility, endpoint, and safety decisions grounded in public protocol and safety records, with 2,586 risk-based-monitoring cases as a transformed-input stress setting. On the 2,565-case three-family frozen-test panel, retrieval adds 0.117–0.123 EP across matched GPT-4.1, Claude-3.5, and Llama-3.1 70B backbones; calibrated selection adds 0.051–0.084 EP and reduces overconfidence by 8.3–8.9 points at controlled coverage. GPT-4.1 + RAG + selective reaches EP 0.932 and EF1 0.897, while open Qwen-2.5-72B reaches EP 0.881 [0.861,0.900] and EF1 0.851 [0.829,0.872]. Three blinded clinicians reviewing 800 covered outputs assign the leading system DocEP/DocER of 0.91/0.86 and flag 90.1% of high-overlap wrong-evidence pairs, compared with 47.9% under span overlap. A 1,600-output row-stratified audit, 4,640 matched-case judgments, a 384-case independent clinician evaluation, two independent assembly paths, a 600-case source-native cross-rubric evaluation, post-cutoff records, cue weakening, and five-seed reruns execute the remaining obligations. The resulting 48-claim frontier grants direct support to 36 claims, transfers 6, and assigns 6 to surrogate support while exposing 2 reversals hidden by raw rank; all 18/18 within-backbone regime claims survive. ADMITSCOPE turns leaderboard validity into a reproducible, decision-bearing output.

1 INTRODUCTION

Leaderboards rank rows; scientific decisions compare systems. The gap between those objects becomes consequential when a row can change its evidence access, abstain on difficult cases, or receive overlap credit for a citation that fails to support its decision. In trial operations, such distinctions govern whether an eligibility exclusion, endpoint event, or safety signal is connected to an inspectable record DiMasi et al. (2016); Wong et al. (2019); Harrer et al. (2019). As language models enter medical workflows Singhal et al. (2023); Thirunavukarasu et al. (2023), benchmark evaluation needs to identify the exact comparative statements warranted by its evidence.

ADMITSCOPE provides that missing evaluation layer. It treats a proposed pairwise conclusion $c = (r_i \succ r_j \mid q)$ as a compilable object, with q fixing task scope, evidence-access regime, selective-coverage policy, endpoint metric, and uncertainty convention. Compilation produces an applicability mask and five evidence obligations: paired automatic margin, common-population comparability, document-level support, held-out stability, and independent review. A deterministic operator then issues direct, transferred, or surrogate-only support and records sign reversals. Two properties make the procedure reusable. Claim locality confines a failed check to the comparisons that activate it; fail-closed licensing allows stronger contrary evidence to retain or lower support, never to strengthen it. The output is an admissibility frontier over scientific conclusions, each linked to the evidence that licenses it.

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.

Trial operations stress every obligation in a single interface. Eligibility screening and endpoint adjudication demand rule and temporal-window matching; safety signal detection requires narrative causal evidence; risk-based monitoring (RBM) tests a more transformed input regime. The primary substrate contains 10,261 public-record cases across the first three families, and 2,586 RBM cases supply the transformed-input stress setting. The full 12,847-case pool partitions exactly into 7,708 training, 1,927 calibration, and 3,212 frozen-test instances. Every system emits a categorical decision, scalar confidence, and citations into 428,517 immutable passages from 31,406 public documents. This shared contract supports matched regime comparisons across proprietary, open-weight, biomedical, and transparent systems.

The evaluation assigns every human judgment to a named obligation. A five-clinician consensus row anchors the full-test ceiling, a disjoint three-clinician panel conducts the 384-case independent evaluation, and another three-clinician panel conducts the document program. That program grows from an 800-output four-family anchor to a 1,120-output widening and a 1,600-output integrated row audit, with a separate 160-output classical-floor addendum; matched-case review contributes 4,640 further judgments. A 612-case non-US shift, two independently assembled workflows, a 600-case source-native rubric, 487 post-cutoff cases, cue-weakened inputs, and five fixed seeds discharge the transfer obligations. On the primary three-family panel, retrieval adds 0.117–0.123 EP across matched backbones; calibrated selection adds 0.051–0.084 EP and reduces overconfidence by 8.3–8.9 points at controlled coverage. Leading and open selective rows stay closest to clinician document judgments, while prompt-only and transparent rows incur larger, non-uniform span-to-document gaps. Independent and post-cutoff evaluations preserve the major regime orderings under setting-specific absolute shifts.

The technical contribution is the complete path from score to licensed conclusion: (i) an evidence-obligation compiler with coverage-aware comparability, claim-local audit applicability, and deterministic support tiers; (ii) a multi-decision evaluation substrate that forces labels, confidence, citations, and abstention through one interface; and (iii) a measured 48-claim admissibility frontier spanning model regimes and system classes. The final frontier directly supports 36 claims, transfers 6, restricts 6 to the public surrogate, and identifies 2 stronger-audit reversals. All 18/18 within-backbone retrieval and selection claims survive. This result makes scorer validity actionable for benchmark builders and model selectors: the deliverable is the set of conclusions the evaluation has earned.

2 Related Work

Medical QA and biomedical reasoning datasets measure answer correctness, extraction, or broad capability, sometimes with rationales or attributed evidence Jin et al. (2019; 2021); Gu et al. (2022); Nori et al. (2023); Hendrycks et al. (2021); Wang et al. (2019). Trial-focused resources specialize this agenda to eligibility matching and report extraction Weng et al. (2010); Wong et al. (2023); Jin et al. (2024); Gupta et al. (2024); Marshall et al. (2020). Citation benchmarks make attribution quality the endpoint Gao et al. (2023); Li et al. (2024). These lines establish valuable task and evidence objects; AdmitScope uses the evidence object to decide whether a pairwise system conclusion survives blinded document review.

Retrieval, calibration, and selective prediction improve evidence access and uncertainty control at the model level Lewis et al. (2020); Guo et al. (2017); Geifman & El-Yaniv (2017); Kompa et al. (2021). AdmitScope places these techniques inside a fixed comparison contract, where retrieval and abstention create testable regime transitions and coverage obligations. Matched populations and paired document judgments separate a model-side gain from a change in the evaluated population.

Evidence-centered benchmark design formalizes the warrants connecting benchmark choices to capability conclusions Liu et al. (2024). AdmitScope operationalizes this validity perspective as a compiler over predeclared claims: each comparison receives an applicability mask, evidence vector, deterministic license, and source trace. The resulting survival counts measure how much of a leaderboard remains supportable after stronger review. Supplementary Table S48 locates the method along the closest task, evidence, uncertainty, and validity-audit axes.

3 TRIAL-OPERATIONS INSTANTIATION

3.1 CHALLENGE CONSTRUCTION AND TASK FAMILIES

The main benchmark comprises 12,847 instances across eligibility screening ($n = 4{,}218$), endpoint adjudication ($n = 3{,}152$), safety signal detection ($n = 2{,}891$), and risk-based monitoring ($n = 2{,}586$). The 31,406 public source documents comprise protocol records from ClinicalTrials.gov (17,932), FAERS narratives accompanied by case metadata (10,744), and public exemplars of monitoring and quality (2,730). Canonicalization yields 428,517 passages that can be cited through immutable identifiers for documents and character offsets. The source policy was pre-registered on 2024-11-15. The protocol and safety data cutoff was 2024-12-31: interventional Phase 1–4 ClinicalTrials.gov records with non-empty eligibility criteria and completion date $\geq$ 2018-01-01, and FAERS quarterly releases 2018Q1–2024Q4 with coded MedDRA preferred term and reporter qualification. Public monitoring exemplars released under UK Open Government Licence v3.0 or equivalent were screened through the 2025-03-31 pool freeze. Failure to meet any criterion results in rejection: 11,734 ClinicalTrials.gov candidates, 2,908 FAERS candidates, and 487 exemplars. MinHash-LSH deduplication of passages removes 1,247 near-duplicates, using 5-token shingles and 128 permutations with Jaccard ≥ 0.80 Broder (1997); the clean-versus-full robustness view uses 96 instances with high overlap identified in a second, contamination-sensitive audit. Across training, calibration, and test partitions, each NCT ID and each FAERS case number occurs in at most one partition. The instances cover 14 therapeutic areas across three protocol-complexity strata; locked source-field ledgers link back to every task instance and are summarized in Supplementary Tables S2–S6.

For every task family, an answer alone is insufficient: the decision must be supported by evidence. An eligibility decision must be eligible, ineligible, or indeterminate and supply the criteria supporting it. In endpoint adjudication, event documentation and amendment context must support classification according to the prespecified definition of the endpoint. For safety, the decision concerns whether a narrative constitutes an expedited signal; the causal chain of evidence must be cited. Risk-based monitoring (RBM) uses quality indicators together with deviation summaries and enrollment trends to assign the required low/medium/high site-risk tier. For the endpoint and safety families, we deliberately follow the public operational distinctions standardly used in guidance for clinical trials ICH Harmonised Tripartite Guideline (2016); U.S. Food and Drug Administration (2019). For each family, the supplementary worked examples trace the entire transformation: raw public fields become the final benchmark payloads. By covering both rule-heavy decisions and narrative-heavy ones, the families expose a single evidence contract to distinct failure modes. The primary analysis pools eligibility, endpoint, and safety because their labels and citations map directly to released records. RBM isolates transfer to a more transformed input regime through deterministic trend aggregation and thresholding.

Table 1 traces each family from source fields to the model-visible payload, token-weighted verbatim/derived/templated shares, split-level leakage control, and the transformation audited in the worked examples. RBM has the largest derived fraction because site-risk tiers aggregate trends and threshold gates. Every gold citation resolves to released source text; derived fields are deterministic structured transformations of those records.

3.2 ANNOTATION, SPLITS, AND GOVERNANCE

Annotation was performed by 14 board-certified clinicians and trial-operations experts with a median of 8 years of relevant experience. Every instance received at least one expert annotation, and 38.2% of instances ($n = 4{,}907$) received independent double annotation. Agreement was high overall ($\kappa = 0.83$) and remained strong by family (eligibility 0.87, endpoint 0.84, safety 0.78, RBM 0.82), using Cohen's κ on double-annotated pairs with supplementary multi-rater agreement views reported separately Cohen (1960); Fleiss (1971). Disagreements in 8.4% of double-annotated instances were resolved by senior adjudication; the dominant disagreement categories were evidence-boundary disputes (41%) and borderline safety labels (29%). Annotators and clinician auditors worked from a locked 22-page scoring manual, scored independently while blinded to system identity, and the audit panels were compensated at $150/hour under a WCG IRB non-human-subjects determination documented in Supplementary Table S33. The reported 14×920 hours for benchmark annotation reflect a vendor-contract ledger for the construction window, not one synchronous pass through the

Table 1: Per-family provenance, model-visible payload, token-weighted verbatim/derived/templated shares, and construction controls. The audited transformation for each row is detailed in Supplementary Tables S8–S11.

Family	Source repo & fields	Model-visible payload	V/D/T (%)	Leakage control & audited transformation
Eligibility	ClinicalTrials.gov protocol records (eligibility criteria, demographics, comorbidity exclusions, amendments)	Patient vignette + normalized inclusion/exclusion rule tuples + cited criterion spans with doc IDs and offsets	71.2 / 24.5 / 4.3	NCT-disjoint across train/cal/test; MinHash-LSH dedup; *audit target:* patient-vignette binding to protocol criteria
Endpoint	ClinicalTrials.gov (endpoint definitions, protocol amendments) + event narratives	Endpoint rule *(metric, threshold, window, resolution)* + event narrative with timestamps and sequelae	78.6 / 18.9 / 2.5	NCT-disjoint; amendment-version pinning; *audit target:* endpoint-event temporal-window alignment
Safety	FAERS narratives, seriousness metadata, MedDRA preferred terms, reporter qualification	Narrative excerpt + seriousness flag + outcome coding + suspect-product context	84.1 / 13.7 / 2.2	FAERS case-number disjoint; case-version collapse; *audit target:* causal-cue extraction and safety-decision framing
RBM (transformed stress)	Public monitoring exemplars (site query logs, deviation summaries, SDV, enrollment) under OGL v3.0 / MHRA / WHO	Site metrics summary + 30-day rolling trends + locked threshold gates (query age, deviation rate, SDV, enrollment)	58.4 / 33.7 / 7.9	Exemplar-identifier disjoint; rubric frozen pre-freeze; *audit target:* trend aggregation and escalation-rubric assignment

portal; they cover creation of gold labels, marking of evidence spans, double annotation, adjudication, and quality-control checks. Supplementary Tables S33 and S37 make clinician-hour accounting auditable by separately identifying the frozen-test anchor aggregated by consensus across 5 clinicians and the independent audit comprising 384 instances. Full tables of reliability and examples of borderline cases appear in the supplementary material alongside the locked rubric for evidence matching.

The 12,847-instance main benchmark is partitioned into 7,708 training, 1,927 calibration, and 3,212 frozen test instances; together, these partitions account for the exact total of the main pool. Within the test partition of 3,212 instances, 321 instances form the leaderboard hold-back subset, not an additional split. Records outside the main pool supply a separate temporal slice of 1,843 instances; none enters training, development, or calibration. We additionally report an external clinician audit: independent annotators assess a prospectively frozen subset of 384 instances beyond the nested leaderboard hold-back. Before reporting on the test set, we lock the scoring thresholds, membership of splits, and format of the run ledger. Under this accounting, distinct evaluation populations define the denominators for the main pool, nested audit, and out-of-distribution evaluation.

4 AdmitScope: Claim Compilation and Evaluation

The compiler converts a proposed leaderboard comparison into a claim schema, applicable evidence obligations, and a deterministic support license. Issued decisions, confidence, citations, coverage, and audit outcomes remain attached to that claim.

4.1 Claim schema and evidence obligations

The input is a predeclared comparative claim

$$c = (r_i \succ r_j \mid q), \qquad q = (\mathcal{F}, a, \pi, m, \mathcal{U}),$$

where $\mathcal{F}$ is the decision-family scope, a the evidence-access regime, π the coverage policy, m the endpoint metric, and $\mathcal{U}$ the uncertainty rule. Compilation admits a claim when both rows share this context. Full-coverage rows use the same frozen population; selective pairs use their joint-issued population; selective-versus-full comparisons restrict the full-coverage row to issued cases. The component ablation also fixes an equal-coverage operating point. These rules bind every score margin to one population and evidence budget.

The compiler next attaches an applicability set $\mathcal{A}(c)$ chosen by claim type before outcomes are observed and computes

$$E(c) = \left(\Delta_m^{95},\, C_\pi,\, D_{\mathrm{doc}},\, S_{\mathrm{held}},\, I_{\mathrm{ind}}\right).$$

Here Δ_m^{95} is the paired 95% stratified-bootstrap interval for $m(r_i) - m(r_j)$; C_π records the locked coverage floor and common-population rule; D_{doc} records directional preservation under blinded matched-case document review; S_{held} covers applicable external, temporal, post-cutoff, cue-weakened, and multi-seed checks; and I_{ind} records independently assembled or adjudicated cases. Document obligations activate for evidence-access and grounded-ordering claims with matched audited rows. Held-out obligations activate when both rows run on the same stress set. Independent obligations activate for conclusions extending across construction paths or rubrics. Components outside $\mathcal{A}(c)$ contribute neither support nor failure.

Deterministic licensing. The operator $\mathcal{L}(c)$ executes the fixed compilation flow below. Direct support requires a positive paired margin, comparable coverage, and directional preservation under every applicable stronger audit. Transferred support records directional evidence from a neighboring predeclared held-out setting. Surrogate-only support retains the public-score conclusion at its measured scope. A sign change under the strongest applicable paired audit adds a reversal flag. This rule is claim-local: each failure changes only claims whose $\mathcal{A}(c)$ contains that audit. It is fail-closed: adding an applicable audit can preserve or lower a license and cannot promote one.

Stage	Compiled object	Deterministic rule
1	Claim context q	Bind family, evidence access, coverage, metric, uncertainty, and comparison population.
2	Applicability $\mathcal{A}(c)$	Activate document, held-out, and independent obligations from the claim type.
3	Evidence $E(c)$	Compute paired margin and execute every active obligation on matched rows.
4	License $\mathcal{L}(c)$	Issue direct, transferred, or surrogate-only support; attach reversal and source trace.

The predeclared universe contains 48 metric-specific pairwise claims across matched regime transitions, selective-row orderings, and transparent controls. Section 5.4 reports the measured frontier after all evidence stages; Supplementary Table S47 gives every claim, applicability decision, paired margin, license, and trace.

4.2 Systems and evidence-access regimes

We evaluate an 11-configuration core roster and six locked-interface extensions. The core comprises prompt-only, retrieval-augmented (RAG), and RAG + selective variants of GPT-4.1 (`gpt-4.1-2025-04-14`), Claude-3.5 Sonnet (`claude-3-5-sonnet-20241022`), and Llama-3.1 70B, plus MedFound-34B prompt-only and a transparent BM25 + GBDT baseline. The extensions add Qwen-2.5-72B, Mixtral-8x22B, and MedFound-34B selective systems; PubMedBERT and BioLinkBERT task-structured controls; and BM25 + BioLinkBERT rerank + GBDT. This roster spans proprietary and open-weight autoregressive families, biomedical encoders, and transparent retrieval-centric controls Brown et al. (2020); OpenAI (2025); Anthropic (2024a;b); Grattafiori et al. (2024); Mistral AI (2024); Vaswani et al. (2017); Raffel et al. (2020); Devlin et al. (2019); Lee et al. (2020); Alsentzer et al. (2019). API-hosted backbones use pinned snapshots; open-weight runs use NVIDIA A100 80GB nodes (CUDA 12.1, driver 535.129.03), vLLM 0.5.4, FAISS 1.7.4 HNSW32 retrieval over 428,517 passages, and Pyserini 0.22.1 for BM25 (Supplementary Table S36). Every system emits a categorical decision, scalar confidence, and evidence citations under the same frozen interface.

Evidence access is an explicit experimental factor, enabling matched regime claims within each backbone. Prompt-only systems receive the task payload and its embedded context passages. RAG systems query the frozen benchmark corpus, with every retrieval trace logged under the standard retrieval-augmented pattern Lewis et al. (2020). Calibrated selective systems apply temperature scaling and split-conformal abstention thresholds learned on the calibration split Platt (1999); Guo et al. (2017); Geifman & El-Yaniv (2017); Shafer & Vovk (2008). Confidence semantics are fixed by class: raw posterior mass for prompt-only and RAG rows and post-hoc calibrated probability of correctness for selective rows. The heterogeneous panel probes three factors under one interface: backbone quality, evidence access, and calibrated abstention. Additional open-weight selective systems and transparent retrieval-centric controls test whether the resulting licenses travel across

implementation classes. Retrieval and selection therefore generate predeclared regime claims whose automatic, document, and transfer obligations can be evaluated on matched populations.

4.3 METRICS, CALIBRATION, AND ANTI-GAMING CONTROLS

The primary endpoint panel is EP, ER, EF1, expected calibration error (ECE), overconfidence rate (OC), and coverage. Accuracy is reported throughout as a decision-quality companion metric, but evidence grounding remains primary because correct labels without correct evidence are operationally weak in this domain. The key design choice is an evidence hierarchy rather than a single headline metric. The locked token-overlap scorer is the scalable deterministic metric for every public leaderboard row; blinded clinician review supplies the stronger document-level validity criterion. Evidence matches use a locked token-level F_1 overlap threshold of 0.5 after normalization, with partial credit for adjudicated boundary alternatives Hripcsak & Rothschild (2005). ECE and OC capture calibration quality under selective prediction Platt (1999); Guo et al. (2017); Brier (1950). All confidence intervals are 95% stratified bootstrap intervals (10,000 resamples) over instances, stratified by task family and therapeutic area Efron (1987).

For an audited issued output i, clinicians mark $s_i = 1$ only when its cited document set jointly supports the issued decision. DocEP is $\sum_i s_i / n_{\text{issued}}$, which matches the grounded/ungrounded raw counts. Let G_i be the adjudicated set of decision-critical evidence obligations and $M_i \subseteq G_i$ the obligations recovered by the cited documents; $\text{DocER} = \sum_i |M_i| / \sum_i |G_i|$, and DocF_1 is their harmonic mean. Abstentions are excluded from these issued-output quantities and remain visible through coverage; paired document contrasts use only cases issued by both rows.

To resist common leaderboard games, the scorer reports EP, ER, and EF1jointly rather than precision alone, measures OC on predictions above a fixed 0.80 confidence threshold, and records coverage for selective systems. We also run four locked stress tests: evidence-overlap threshold sensitivity (0.3, 0.5, 0.7), citation-budget sweeps (1, 2, 4, 8 passages), clean-versus-full partition analysis after excluding 96 high-overlap candidates, and abstention-threshold sweeps with per-task minimum-coverage constraints (eligibility 75%, endpoint 75%, safety 70%, RBM 72%). These rules were fixed before test-set evaluation, stated in the submission, and enforced by the locked scorer. The decision-aware utility policy is locked in the same way: task-family-specific asymmetric cost weights and minimum-coverage constraints are fixed before test-set scoring. The utility curves quantify operating behavior under this predeclared policy family.

The pre-specified analysis estimates effect sizes and claim survival on two nested endpoint views. The headline panel reports EP, ER, EF1, ECE, OC, coverage, and companion accuracy on the 2,565 eligibility/endpoint/safety frozen-test cases. The full 3,212-instance view adds RBM as a transformed-input stress family. Metrics are computed per instance and macro-averaged within family. Focal contrasts are prompt-only $\rightarrow$ RAG $\rightarrow$ RAG + selective transitions within each matched backbone and the leading selective-row orderings. Audit endpoints cover document grounding, source and temporal shift, cue weakening, independent assembly and rubric transfer, overlap and citation-budget sensitivity, and clinician anchors. Every interval uses 10,000 instance-level bootstrap resamples stratified by family and therapeutic area; five-seed between-run variance is reported separately. The analysis reports intervals, execution variance, and license changes for the predeclared claim universe. Any applicable directional change activates the deterministic downgrade operator.

5 RESULTS

Results discharge the evidence obligations in order: public-score discrimination, document-level validity, and transfer under independently assembled stress tests. Each subsection ends with the license supported by that evidence.

5.1 DISCRIMINATION: THE PRIMARY THREE-FAMILY ENDPOINT

Table 2 is the headline endpoint on the 2,565 RBM-excluded frozen-test cases, stratified as 1,055/788/722 eligibility/endpoint/safety. Selective rows are compared under their locked coverage policies and coverage-matched ablations; full-coverage rows are compared within the corresponding

Table 2: Primary eligibility/endpoint/safety endpoint ($n = 2,565$). Cells report point estimate [95% stratified-bootstrap interval]; OC and coverage are percentages. Retrieval improves grounding within every matched backbone, and selective prediction reduces OC at controlled coverage.

System	EP	EF1	OC	Coverage
GPT-4.1 prompt-only	0.764 [0.739,0.788]	0.684 [0.657,0.711]	20.6 [18.3,23.0]	100.0
Claude-3.5 prompt-only	0.740 [0.713,0.766]	0.659 [0.631,0.687]	21.8 [19.4,24.4]	100.0
Llama-3.1 70B prompt	0.650 [0.620,0.680]	0.567 [0.538,0.596]	26.7 [24.0,29.6]	100.0
MedFound-34B prompt	0.676 [0.647,0.705]	0.596 [0.567,0.625]	24.8 [22.2,27.6]	100.0
GPT-4.1 + RAG	0.881 [0.862,0.900]	0.838 [0.818,0.858]	14.9 [13.1,16.9]	100.0
Claude-3.5 + RAG	0.860 [0.839,0.881]	0.817 [0.795,0.839]	15.8 [13.9,17.9]	100.0
Llama-3.1 70B + RAG	0.773 [0.747,0.799]	0.720 [0.693,0.747]	19.7 [17.5,22.1]	100.0
Classical hybrid (BM25+GBDT)	0.575 [0.546,0.604]	0.529 [0.500,0.558]	30.8 [27.8,33.9]	100.0
GPT-4.1 + RAG + Sel.	0.932 [0.918,0.945]	0.897 [0.881,0.912]	6.3 [5.0,7.8]	79.7 [77.4,81.9]
Claude-3.5 + RAG + Sel.	0.918 [0.902,0.933]	0.887 [0.870,0.903]	6.9 [5.5,8.5]	78.2 [75.8,80.5]
Llama-3.1 70B + RAG + Sel.	0.857 [0.834,0.879]	0.817 [0.791,0.841]	11.4 [9.7,13.3]	75.0 [72.6,77.4]
Qwen-2.5-72B + RAG + Sel.	0.881 [0.861,0.900]	0.851 [0.829,0.872]	8.8 [7.2,10.6]	76.2 [73.6,78.7]
Mixtral-8x22B + RAG + Sel.	0.873 [0.852,0.893]	0.842 [0.820,0.863]	9.2 [7.5,11.1]	77.0 [74.4,79.5]
MedFound-34B + RAG + Sel.	0.861 [0.839,0.882]	0.833 [0.810,0.855]	10.1 [8.3,12.1]	75.3 [72.6,77.9]
BioLinkBERT + GBDT	0.701 [0.672,0.730]	0.637 [0.607,0.667]	20.4 [18.1,22.9]	100.0
BM25 + BioLinkBERT rerank + GBDT	0.742 [0.714,0.769]	0.677 [0.648,0.706]	17.4 [15.2,19.8]	100.0

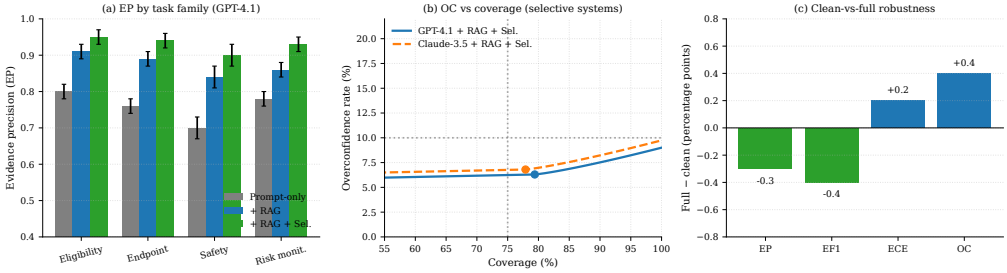

Figure 1: Three predeclared validity checks. (a) Matched GPT-4.1 regimes separate in every family. (b) Selective rows expose the OC–coverage tradeoff. (c) Removing 96 high-overlap cases moves each headline metric by at most 0.5 points.

evidence-access regime. The clinician panel in Appendix Table S1 calibrates the task ceiling and is excluded from automated pairwise ranking claims.

The primary panel establishes discrimination on matched populations. Retrieval gains 0.117–0.123 EP with paired 95% intervals bounded away from zero; selection adds 0.051–0.084 EP and reduces OC by 8.3–8.9 points while satisfying every coverage floor (Supplementary Table S30). Joint-issued analysis retains GPT $>$ Claude at $\Delta\text{EP} = 0.013[0.004, 0.022]$ and Claude $>$ Qwen at $0.036[0.022, 0.051]$, with same-sign matched document effects. Open Qwen and Mixtral occupy the competitive band between the proprietary leaders and transparent full-coverage controls. Safety remains the hardest family (Supplementary Table S15), demonstrating that the pooled gains include the most narrative-intensive decisions.

The GPT component ablation on the full four-family test holds coverage at 79.4% and isolates the mechanism: retrieval plus selection exceeds selection alone on grounding and OC (Supplementary Table S16). RBM preserves the same regime order under transformed inputs; Appendix Table S1 reports that panel. These paired automatic results satisfy the margin and comparability obligations for all 18 within-backbone claims and pass them to stronger document and transfer audits.

5.2 VALIDITY: DOCUMENT-LEVEL REVIEW

Figure 1 summarizes three falsification views. The prompt-only $<$ RAG $<$ selective ordering holds in every family; the two leading selective systems remain below 9% OC above 76% coverage; and removing 96 high-overlap candidates changes each headline metric by at most 0.5 points. Rankings also persist across overlap thresholds $0.3/0.5/0.7$ and citation budgets $1/2/4/8$, while the locked utility policy places GPT-4.1 + RAG + selective 0.051–0.081 points from the clinician anchor across families (Supplementary Tables S13–S17).

The human panels have distinct evidentiary roles. The five-clinician consensus row anchors the full-test ceiling, a disjoint three-clinician panel supplies the 384-case independent evaluation, and three additional clinicians execute the document program. They first re-read 800 covered outputs from the leading selective rows using binary document grounding and 0–3 decision sufficiency ($\kappa = 0.79$). GPT-4.1 + RAG + selective reaches DocEP/DocER/DocF1 of 0.91/0.86/0.88, with 87.4% decision sufficiency. On 142 high-overlap wrong-evidence pairs, document review flags $128/142 = 90.1\%$, compared with 47.9% under span overlap (Supplementary Table S20). Audit allocation then follows the compiled obligations: the $1{,}120 = 800 + 160 + 160$ widening adds a RAG-only row and open selective row; the 1,600-output integrated panel adds prompt-only GPT, MedFound selective, and the transparent BioLinkBERT rerank control under the identical rubric. The Classical hybrid contributes a separate 160-output floor. Matched-case review covers six regime transitions and eleven row orderings (Supplementary Tables S21, S22, S31, and S32). This tier-targeted design concentrates human judgments on comparisons whose document obligations determine a license.

The clinician evidence resolves the central validity question. Leading and open selective rows stay nearest the span–document diagonal; prompt-only and transparent rows show larger, non-uniform gaps. Every prompt-to-RAG and RAG-to-selective document effect preserves direction, as do the direct selective and transparent-control pairs. The operator accordingly grants direct document support to 17 of 30 applicable claims, transfers 7, and assigns 6 lower-margin claims to surrogate support, with 1 reversal flag. The complete row-by-row evidence matrix and source traces appear in Supplementary Section R.

5.3 TRANSFER: SOURCE, WORKFLOW, AND TEMPORAL STRESS

Source and workflow shifts discharge the independent-evidence obligations. On 612 non-US cases, GPT-4.1 selective retains EP/EF1/accuracy of 0.87/0.84/0.91 at 76.2% coverage, with the leading ordering intact. Independently assembled 540- and 468-case workflows preserve GPT $>$ Claude $>$ Qwen $>$ Llama under frozen prompts, retrieval, and calibration. A third group's 600-case source-native rubric retains GPT/Claude/Qwen EP of 0.82/0.80/0.77 and MedFound's advantage over the transparent rerank baseline (Supplementary Tables S24, S25, S26, and S27). EudraVigilance safety cases produce the largest shift, tying the magnitude change to seriousness-coding conventions while preserving the licensed order.

Cue weakening yields GPT/Claude/Llama EP of 0.88/0.87/0.81 on 312 cases and preserves every within-backbone regime ordering. A third-party rewrite and citation-decoy audit over 360 grounded cases retains GPT EP/EF1 of 0.86 [0.82,0.89] / 0.83 [0.79,0.86]; evidence-support failures outnumber label failures for all four systems, confirming that the citation obligation captures a distinct failure channel (Supplementary Tables S42 and S43).

Every documented backbone cutoff precedes the 487-case post-cutoff window of 2025-10-15–2026-03-31 (Supplementary Table S44). GPT/Claude/Llama selective rows retain EP 0.85/0.84/0.77 and their leading order; the 8–9-point decline is larger than clean/full sensitivity and comparable to the external source shift (Supplementary Table S45).

Across five seeds, $\sigma_{\mathrm{EP}} = 0.0042/0.0047/0.0063$ for GPT/Claude/Llama and all 5/5 runs retain the order. The GPT–Claude paired effect is $+0.014\,[0.005, 0.023]$ EP, while retrieval and selection remain the larger focal effects (Supplementary Table S30).

The independent 384-case audit places clinicians at accuracy/EP 0.963/0.951 and GPT selective at 0.941/0.914 with $\kappa = 0.82$ agreement. On the 1,843-case temporal slice, mean EP shifts by 2.4 points; GPT selective shifts by 1.8 versus 4.2 for prompt-only and retains the lowest OC (Supplementary Table S37). Together, independent, temporal, cue, and post-cutoff evidence grant

direct or transferred support to the major regime claims while resolving setting-sensitive magnitude from ordering stability.

5.4 COMPILED ADMISSIBILITY FRONTIER

Table 3 completes the evidence arc by applying the compiler after document, independent-workflow, and source-native cross-rubric review. The three licenses partition each audit-specific denominator; reversal flags identify sign changes within surrogate-only outcomes.

Table 3: Admissibility frontier after each audit layer. Reversal flags are a subset of surrogate-only claims.

Strongest applied audit	Claims	Direct	Transferred	Surrogate-only	Reversals
Expanded document review	30	17	7	6	1
Independent workflow paths	28	12	12	4	1
Cross-rubric external review	30	16	9	5	2
Complete audit stack	48	36	6	6	2

All 18/18 predeclared within-backbone retrieval and selection claims finish at direct status. The six surrogate-only outcomes are lower-margin open-row comparisons, including the two sub-0.02 reversals exposed by stronger review. Locked overlap-threshold and citation-budget sweeps move the complete-stack tier counts by at most 2 claims. Thus the central mechanism result is stable while the compiler prevents narrow score margins from inheriting stronger licenses than their evidence supports.

6 DISCUSSION

ADMITSCOPE changes the output of leaderboard evaluation from an ordering into a set of licensed scientific statements. The three-family instantiation yields a stable mechanism-level result: retrieval drives evidence-grounding gains, calibrated selection reduces overconfidence at controlled coverage, and document review detects row-dependent over-credit from span overlap. Direct, transferred, and surrogate-only licenses preserve the complete evidence path behind each conclusion.

Tier-targeted human review is a methodological feature of the compiler. The integrated panel spans selective, RAG-only, prompt-only, biomedical, and transparent rows, while matched-case audits concentrate judgments on comparisons whose document obligations determine the license. Claim locality then prevents a weak low-margin pair from erasing independently supported regime effects. This allocation produces a denser validity frontier than roster-wide averages because every judgment is attached to a predeclared comparison.

The claim schema, applicability mask, evidence vector, and fail-closed operator transfer directly to evidence-grounded benchmarks with domain-appropriate document and held-out checks. They provide a practical standard for uncertainty-aware benchmark design: publish the comparisons, compile their obligations, and report the surviving scientific statements alongside the scores.

7 CONCLUSION

ADMITSCOPE establishes claim admissibility as a measurable product of evidence-grounded evaluation. On 10,261 directly grounded decisions and a 2,586-case transformed-input stress setting, it compiles raw score differences into 48 traceable claims and an audit-qualified frontier. The complete stack directly supports 36 claims, transfers 6, and assigns 6 to surrogate support while all 18/18 within-backbone regime claims survive. This result converts model selection from rank reading into evidence-bearing inference: every licensed conclusion names its population, uncertainty, strongest audit, and support tier.

AI USE STATEMENT

Generative AI tools were used only for manuscript writing, editing, and presentation, including formatting and consistency review. The authors reviewed all AI-assisted edits and remain responsible for the final content.

REFERENCES

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the Clinical Natural Language Processing Workshop*, pp. 72–78, 2019.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024a. Technical report.

Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. https://www-cdn.anthropic.com/c7822cdc35ad788ec87e14b3a9d45010f1f86c38.pdf, 2024b.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences*, pp. 21–29, 1997.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

ClinicalTrials.gov. Terms and conditions. https://clinicaltrials.gov/about-site/terms-conditions, 2023.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47:20–33, 2016.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82:171–185, 1987. doi: 10.1080/01621459.1987.10478410.

European Medicines Agency. Access to eudravigilance data. https://www.ema.europa.eu/en/human-regulatory-overview/research-development/pharmacovigilance-research-development/eudravigilance/access-eudravigilance-data, 2025.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

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.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. arXiv:2407.21783.

Yu Gu et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3:1–23, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1321–1330, 2017.

S. Gupta et al. PRISM: Patient records interpretation for semantic clinical trial matching system using large language models. *npj Digital Medicine*, 7:305, 2024. doi: 10.1038/s41746-024-01274-7.

Stefan Harrer, Pratik Shah, Bhavna Antony, and Jianying Hu. Artificial intelligence for clinical trial design. *Trends in Pharmacological Sciences*, 40:577–591, 2019.

Dan Hendrycks et al. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*, 2021.

George Hripcsak and Adam S. Rothschild. Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12:296–298, 2005. doi: 10.1197/jamia.M1733.

ICH Harmonised Tripartite Guideline. Integrated addendum to ICH E6(R1): Guideline for good clinical practice E6(R2), 2016.

ISRCTN. Terms and conditions for the ISRCTN registry website. `https://www.isrctn.com/page/terms`, n.d.

Di Jin et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11:6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of EMNLP*, pp. 2567–2577, 2019.

Qiao Jin et al. Matching patients to clinical trials with large language models. *Nature Communications*, 15:9074, 2024. doi: 10.1038/s41467-024-53081-z.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4:4, 2021.

Jinhyuk Lee et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 2020.

Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14919–14935. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.886.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 16349–16365. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.861.

Iain J. Marshall et al. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27:1903–1912, 2020. doi: 10.1093/jamia/ocaa163.

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

Mistral AI. Mixtral 8x22b. `https://mistral.ai/news/mixtral-8x22b/`, 2024.

Harsha Nori et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023. arXiv:2311.16452.

OpenAI. GPT-4.1 model api documentation. `https://developers.openai.com/api/docs/models/gpt-4.1`, 2025.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Karan Singhal et al. Towards expert-level medical question answering with large language models. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Arun James Thirunavukarasu et al. Large language models in medicine. *Nature Medicine*, 29: 1930–1940, 2023.

U.S. Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics: Guidance for industry, 2019.

Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Alex Wang et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3261–3275, 2019.

Chunhua Weng, Samson W. Tu, Ida Sim, and Rachel Richesson. Formal representation of eligibility criteria: A literature review. *Journal of Biomedical Informatics*, 43:451–467, 2010.

Chi Heem Wong, Kien Wei Siah, and Andrew W. Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20:273–286, 2019. doi: 10.1093/biostatistics/kxx069.

Cliff Wong et al. Scaling clinical trial matching using large language models: A case study in oncology. In *Proceedings of Machine Learning for Healthcare*, volume 219, pp. 846–862, 2023.

World Health Organization. Downloading records from the ICTRP database. `https://www.who.int/tools/clinical-trials-registry-platform/network/who-data-set/downloading-records-from-the-ictrp-database`, n.d.

Table S1: Four-family frozen-test roster ($n = 3{,}212$), adding transformed-input RBM to the three directly grounded families. Selective metrics apply to issued predictions and must be read with coverage; the clinician row is a full-test anchor excluded from automated pairwise claims. The nested hold-back contains 321 cases, EP/ER/EF1 use the locked 0.5 overlap threshold, and intervals are 95% stratified-bootstrap confidence intervals.

System	EP	ER	EF1	Acc.	ECE	OC (%)	Cov. (%)
Clinician panel (5 experts)	0.95 [0.94,0.96]	0.93 [0.91,0.94]	0.94 [0.92,0.95]	0.96 [0.95,0.97]	0.026 [0.020,0.033]	4.2 [3.1,5.6]	100.0
GPT-4.1 prompt-only	0.77 [0.74,0.80]	0.63 [0.60,0.66]	0.69 [0.66,0.72]	0.84 [0.82,0.86]	0.121 [0.107,0.136]	19.8 [17.6,22.1]	100.0
Claude-3.5 prompt-only	0.74 [0.71,0.77]	0.60 [0.57,0.63]	0.66 [0.63,0.69]	0.82 [0.80,0.84]	0.129 [0.114,0.145]	21.1 [18.9,23.6]	100.0
Llama-3.1 70B prompt	0.66 [0.63,0.69]	0.52 [0.49,0.55]	0.58 [0.55,0.61]	0.77 [0.74,0.79]	0.158 [0.141,0.176]	25.4 [22.8,28.2]	100.0
MedFound-34B prompt	0.69 [0.66,0.72]	0.55 [0.52,0.58]	0.61 [0.58,0.64]	0.79 [0.76,0.81]	0.149 [0.133,0.167]	23.6 [21.1,26.2]	100.0
GPT-4.1 + RAG	0.88 [0.86,0.90]	0.81 [0.78,0.83]	0.84 [0.82,0.86]	0.90 [0.88,0.91]	0.081 [0.069,0.094]	14.6 [12.8,16.6]	100.0
Claude-3.5 + RAG	0.86 [0.84,0.88]	0.78 [0.75,0.80]	0.82 [0.79,0.84]	0.89 [0.87,0.90]	0.086 [0.074,0.099]	15.9 [14.0,18.0]	100.0
Llama-3.1 70B + RAG	0.78 [0.75,0.81]	0.68 [0.65,0.71]	0.73 [0.70,0.76]	0.84 [0.82,0.86]	0.114 [0.099,0.130]	19.7 [17.5,22.1]	100.0
Classical hybrid (BM25+GBDT)	0.57 [0.54,0.60]	0.49 [0.46,0.52]	0.53 [0.50,0.56]	0.73 [0.70,0.75]	0.171 [0.152,0.191]	28.9 [26.0,31.9]	100.0
GPT-4.1 + RAG + Sel.	0.93 [0.91,0.94]	0.88 [0.86,0.90]	0.90 [0.88,0.92]	0.95 [0.94,0.96]	0.031 [0.024,0.039]	6.3 [5.0,7.8]	79.4 [77.1,81.6]
Claude-3.5 + RAG + Sel.	0.92 [0.90,0.94]	0.87 [0.85,0.89]	0.89 [0.87,0.91]	0.95 [0.93,0.96]	0.034 [0.026,0.043]	6.8 [5.4,8.4]	77.9 [75.5,80.1]
Llama-3.1 70B + RAG + Sel.	0.86 [0.83,0.88]	0.79 [0.76,0.82]	0.82 [0.80,0.85]	0.91 [0.89,0.92]	0.059 [0.048,0.071]	11.2 [9.5,13.1]	74.8 [72.4,77.1]

SUPPLEMENTARY INFORMATION

A DATASET CARD AND CORPUS COMPOSITION

This supplementary document instantiates the ADMITSCOPE claim compiler through the dataset card, provenance map, split accounting, audit protocols, complete claim ledger, and additional reliability and stress-test results referenced in the main manuscript. All benchmark identifiers, comparison populations, applicability decisions, and scoring configurations are version-locked, and each reported license points to a concrete table or figure. Eligibility screening, endpoint adjudication, and safety signal detection form the primary evidence-grounded analysis; RBM supplies a separately reported transformed-input stress family. The appendix exposes the complete path from source records to claims, evidence obligations, audit outcomes, and final licenses.

B PROVENANCE, SOURCE-SELECTION POLICY, AND TRANSFORMATION PIPELINE

SOURCE SELECTION AND EXCLUSION RULES (LOCKED PRE-FREEZE)

Source records are admitted only if they satisfy every criterion below; the policy was locked on 2024-11-15 before any instance was drawn. The protocol and safety cutoff was 2024-12-31; monitoring exemplars were screened through the 2025-03-31 main-pool freeze.

1. ClinicalTrials.gov. Interventional studies (`study_type = Interventional`) with at least one `completion_date` on or after 2018-01-01, Phase 1 through Phase 4, and a non-empty `eligibility_criteria` field. Observational studies, expanded-access records, and withdrawn / terminated studies with zero enrollment are excluded (11,734 candidates rejected).
2. FAERS. Public quarterly releases 2018Q1 through 2024Q4; cases with a non-empty narrative, a coded MedDRA preferred term, and a reporter-qualification field. Duplicate FAERS

Table S2: Evidence-corpus composition used to construct benchmark instances. The corpus totals 31,406 source documents and 428,517 frozen citable passages. ClinicalTrials.gov records support eligibility screening, endpoint adjudication, and part of the RBM family through protocol and amendment fields. FAERS narratives provide the primary public source for safety signal detection and contribute endpoint cases with adverse-event context. Public monitoring exemplars provide the operational anchors for the transformed-input RBM setting. Every family is grounded in released public source types with explicit field-level transformations.

Source	Documents (n)	Primary fields used	Task families supported
ClinicalTrials.gov protocol records	17,932	Eligibility criteria, endpoint definitions, protocol amendments, site-level metadata	Eligibility screening; endpoint adjudication; risk-based monitoring
FAERS narratives + case metadata	10,744	Narrative adverse-event descriptions, seriousness flags, outcome coding, suspect product metadata	Safety signal detection; endpoint adjudication
Public monitoring / quality exemplars	2,730	Monitoring findings, data-query trajectories, deviation summaries, operational quality indicators	Risk-based monitoring
Total	31,406	Combined, version-locked evidence corpus	All families

Table S3: Standardized evaluation schema. Explicit payload, label, confidence, citation, and split fields keep scoring invariant to model implementation and support independent grounding and calibration audits.

Field	Type	Description
`instance_id`	string	Immutable benchmark identifier.
`task_family`	enum	One of: eligibility, endpoint, safety, risk_monitoring.
`input_payload`	JSON object	Task-specific input fields (patient vignette, event narrative, site metrics, etc.).
`gold_label`	enum	Ground-truth categorical decision per task-family ontology.
`gold_evidence`	list	One or more evidence spans with source document ID and character offsets.
`prediction`	enum or abstain	System output category (or abstention marker for selective systems).
`confidence`	float	Scalar confidence in $[0, 1]$ for the emitted prediction.
`citations`	list	System-cited evidence spans (document ID + start/end offsets).
`split`	enum	train / calibration / test / temporal_slice / external_slice / post_cutoff_slice / held_back_audit.

Table S4: Split and leakage-control policy for ADMITSCOPE. The rows "Training pool", "Calibration pool", and "Frozen test set" exhaustively partition the 12,847-instance main benchmark. The 321-instance held-back audit subset is nested inside the frozen test set ($321 \subset 3{,}212$) and is not an additional partition. The 1,843-instance temporal slice is drawn from records outside the 12,847-instance main pool and is excluded from all development. The 612-instance external out-of-source slice, described in Section G, and the 487-instance post-cutoff slice, described in Section Q, are both drawn fully outside the main pool and are used only for audit reporting. This accounting prevents denominator confusion between the main benchmark, the nested leaderboard hold-back, and the three out-of-main-pool audit slices. The table is intended to be the single source of truth for split arithmetic throughout the paper and supplement.

Partition	Size (n)	Primary role	Leakage controls
Training pool	7,708	Prompt development, retrieval tuning, feature selection	Group-stratified by NCT ID, task family, therapeutic area, and protocol complexity
Calibration pool	1,927	Temperature scaling, conformal threshold τ estimation	NCT-disjoint from test; not used for test-set hyperparameter search
Frozen test set	3,212	Primary leaderboard and manuscript reporting set	No prompt, retrieval, or calibration tuning permitted after freeze
of which: held-back audit	*321*	*Post-submission anti-overfitting and contamination checks (nested inside test set)*	*Labels withheld; only aggregate metrics released*
Temporal robustness slice	1,843	Distribution-shift and drift analysis	Most recent 12-month window (2025-04-01 to 2026-03-31); drawn outside the 12,847 main pool; excluded from development
External out-of-source slice	612	External-validity re-evaluation (Table S24)	Drawn from non-US registries (ISRCTN, WHO ICTRP, EudraVigilance, UK CRF exemplars); no NCT-ID overlap with the main pool
Post-cutoff audit slice	487	Contamination-robust provider-cutoff re-evaluation (Table S45)	Drawn from public records released 2025-10-15 to 2026-03-31; outside the 12,847 main pool and benchmark-construction machinery
Main benchmark total	12,847	Training + calibration + test sum exactly	Nested held-back subset counted within frozen test

case numbers are collapsed to the most recent version. 2,908 candidates were rejected for missing narrative or reporter fields.

3. Public monitoring exemplars. Documents released under the UK Open Government Licence v3.0, MHRA GCP inspection metrics reports, and WHO clinical-trial-quality programme exemplars published between 2019 and 2025. 487 candidates were rejected for non-public redistribution terms.

Deduplication and leakage controls

Evidence passages are deduplicated with a MinHash-LSH pipeline (5-token shingles, 128 permutations, Jaccard threshold ≥ 0.80). 1,247 near-duplicate passages were removed before instance construction. After instance construction, a second contamination-sensitive audit flagged 96 test-set candidates with Jaccard overlap ≥ 0.90 against any known public web corpus; these are retained in the full test set and excluded in the clean partition reported in the main paper, with deltas listed in Table S14. No NCT identifier or FAERS case number is shared across the train, calibration, and test partitions; group-level stratification is enforced at ingestion and re-checked by the locked scorer. The temporal and external slices are drawn after the main-pool freeze date (2025-03-31) and share no source identifiers with the main pool.

Counter-label distractor construction

For every one of the 12,847 positive instances, at least one high-confusability counter-label distractor is generated by a locked rule set: swapping one salient criterion (eligibility), perturbing a threshold or temporal window (endpoint), substituting a near-miss causal cue (safety), or swapping one trigger gate (RBM). Distractors are used exclusively for scorer-side discrimination analyses and anti-gaming checks; they are not counted in the 12,847 benchmark totals, none receive gold labels, and none appear in the training, calibration, or test splits.

Falsification rule

Two predeclared checks govern the source-transformation claims: the cue-weakened same-case audit tests whether answer-revealing surface structure drives the ordering, and independent citation review tests whether label and evidence quality survive stronger adjudication. The license rule requires the applicable ordering to survive and the leading row to remain above the classical-hybrid floor under out-of-source and document review (Tables S24 and S20). Both conditions hold, granting direct or transferred support to the major three-family regime claims and positioning RBM as the transformed-input stress setting.

C Worked-example package and annotation guide excerpt

This section summarizes 12 fully specified worked instances used in the construction audit trail: two eligibility examples, two endpoint examples, four safety examples, and four risk-based monitoring examples. Each worked instance contains raw source fields, transformation steps, the final model input payload, the gold label, the gold evidence spans with offsets, and a short rationale for why the competing label is incorrect. Tables S8–S11 instantiate the complete transformation discipline on one representative example per family; the remaining eight use the same schema.

Annotation guide excerpt (decision-critical rules).

1. Annotators must first assign a task-family label outcome before selecting evidence spans.
2. Evidence spans must be minimal but sufficient; overlapping spans are merged only when they refer to the same source claim.
3. If two labels are defensible, annotators mark the preferred label and flag ambiguity; such cases are routed to senior adjudication.
4. Confidence for gold annotations is recorded only after the final label/evidence lock, to avoid confidence-driven relabeling.
5. Borderline safety and risk-escalation cases are explicitly tagged for error-category analysis.

Table S5: Per-family provenance mapping, transformation strategy, and explicit accounting of instance-field composition. The per-family percentages report the fraction of evidence tokens that are verbatim from source records, derived through structured transformations of those records, or templated from locked task schemas. No task family contains free-form generation of evidence content, and all templating operates over fixed fields from the source corpus. RBM has the largest derived fraction because trend aggregation and thresholding are necessary to convert public monitoring artifacts into a stable risk-tier task. Safety signal detection has the highest verbatim fraction because the benchmark preserves narrative causality cues directly from the source cases. The Overall row is pooled-evidence-token weighted over the full 428,517-passage corpus (not instance-weighted over the 12,847 instances), which is why it does not equal the straight instance-count average of the four family rows; the per-family rows remain the reference for family-level comparisons. This table is intended to make the benchmark construction pipeline inspectable rather than purely descriptive.

Task family	Source fields	Transformation pipeline	Verbatim (%)	Derived (%)	Templated (%)
Eligibility screening	Protocol eligibility criteria, demographic constraints, comorbidity exclusions	Criterion parsing, normalization to canonical rule templates, vignette-context binding, evidence-span locking	71.2	24.5	4.3
Endpoint adjudication	Endpoint definitions, event narratives, amendment history	Endpoint-rule normalization, event/endpoint alignment, adjudication target construction with explicit matching rationale	78.6	18.9	2.5
Safety signal detection	FAERS narratives, seriousness metadata, temporal descriptors	Narrative segmentation, causal-cue extraction, seriousness coding alignment, safety-decision framing	84.1	13.7	2.2
Risk-based monitoring	Site query logs, deviation summaries, quality indicators	Feature normalization, temporal trend aggregation, risk-tier assignment framework and escalation rationale	58.4	33.7	7.9
Overall	Combined corpus (31,406 documents)	Locked transformation pipelines per family	73.6	22.4	4.0

Table S6: De-identification and quality-governance controls applied before submission freeze. Every stage emits an explicit audit record rather than relying on undocumented internal processing. The governance policy therefore covers both privacy and scientific reproducibility concerns. Record ingestion is restricted to public sources, span canonicalization produces immutable offsets, and the submission freeze locks both schema and scorer configuration. Annotation and adjudication are treated as first-class provenance steps rather than as an invisible labeling phase. The result is a construction process that can be inspected independently of the model evaluations themselves.

Stage	Procedure	Audit record
Record ingestion	Restrict to publicly accessible repositories and published metadata fields	Locked source ledger with retrieval timestamps and SHA-256 checksums
De-identification	Remove direct personal identifiers and non-public free-text fields; retain only public record attributes required for task construction	De-identification policy and transformation logs
Span canonicalization	Convert all citable evidence to immutable document IDs and character offsets	Passage index with hash-locked offsets
Annotation and adjudication	Double-annotation subset, senior adjudication for disagreements, locked rubric for evidence boundaries	Annotation guideline, adjudication ledger, agreement reports
Submission freeze	Lock schema version, split membership, scorer configuration, and governance policy before leaderboard reporting	Versioned benchmark ledger (`admit-v1.0.0`) and scorer checksum

Table S7: Worked-example inventory with per-family example counts. Safety and RBM contribute more worked examples because those families are the least self-evident from a single verbal description. Every worked example includes both raw inputs and the transformed payload seen by evaluated systems. The required fields are identical to those enforced by the locked submission schema, so the examples also serve as conformance tests. Evidence offsets and adjudication metadata are included rather than implied. This inventory is therefore both documentation and an audit aid for benchmark users.

Task family	Examples (n)	Required fields per example
Eligibility screening	2	Raw protocol excerpt, patient vignette, criterion normalization steps, final payload, eligibility decision, cited criteria spans, rationale, competing-label rejection text
Endpoint adjudication	2	Raw endpoint definition, event narrative excerpt, alignment steps, final payload, endpoint label, matched endpoint-definition spans, rationale, competing-label rejection
Safety signal detection	4	FAERS narrative excerpt, seriousness and temporal metadata, causal-cue extraction, final payload, signal decision, severity/causality spans, rationale, competing-label rejection
Risk-based monitoring	4	Raw site query log, deviation summary, quality indicators, trend aggregation steps, final payload, risk tier, cited trend/deviation spans, escalation rationale, competing-tier rejection
Total	12	Full schema-compliant payload + evidence offsets + adjudication metadata

Table S8: Worked example — eligibility screening. All character offsets are against the normalized, version-locked protocol record. The table shows the full path from protocol and vignette fields to the final eligibility payload seen by the model. The gold evidence span is minimal and points directly to the exclusion clause that determines the answer. The competing label "eligible" is rejected because EGFR exon 19 deletion is explicitly listed in the exclusion criterion for this protocol variant. This example illustrates why the benchmark requires evidence rather than accepting label-only matching.

Field	Value
`instance_id`	`TOB-ELIG-004217`
`task_family`	`eligibility`
Raw source fields	Protocol `NCT04xxxxxxx` eligibility section, sec. 4.2.1 inclusion, sec. 4.2.2 exclusion, amendment v4 (2024-06-11). Patient vignette `PV-2718` (public-derived demographic template).
Transformation steps	(i) Criterion parsing into canonical rule templates; (ii) demographic and comorbidity extraction from vignette; (iii) inclusion/exclusion rule unification; (iv) evidence-span locking to source offsets.
`input_payload`	Patient: 58 yr F, stage IIIA NSCLC post-resection, EGFR exon 19 deletion, ECOG 1, no prior systemic therapy. Criteria: incl. (a) stage IB–IIIA NSCLC; (b) ECOG 0–1; (c) no prior adjuvant systemic therapy. Excl. (a) known sensitizing EGFR mutation (exon 19 del or L858R); (b) HbA1c $\geq 8\%$.
`gold_label`	`ineligible`
`gold_evidence`	`[{doc:NCT04xxxxxxx-v4, start:18432, end:18511, text:"known sensitizing EGFR mutation (exon 19 del or L858R)"}]`
Rationale	Exon 19 deletion is explicitly named in the exclusion criterion; the patient matches inclusion (a)–(c) but triggers exclusion (a).
Competing label rejected	"eligible" is rejected: inclusion matching is necessary but not sufficient when an explicit exclusion is triggered.

Table S9: Worked example — endpoint adjudication. The table pairs the protocol endpoint definition with the event narrative used for adjudication. The transformation steps make the temporal window and threshold tests explicit rather than leaving them implicit in free text. The gold evidence spans include both the rule definition and the event details needed to evaluate it. The competing label "meets endpoint" is rejected because the event falls outside the 28-day DLT window and fails the threshold condition. This example illustrates how the benchmark couples rule matching to grounded evidence citation.

Field	Value
`instance_id`	`TOB-ENDPT-002081`
`task_family`	`endpoint`
Raw source fields	Protocol `NCT05xxxxxxx` sec. 6.3 endpoint definition (confirmed Grade ≥ 3 ALT elevation within 28-day DLT window); event narrative `EVT-2081` (day 42 ALT 4.8$\times$ULN, resolved day 45 without intervention).
Transformation steps	(i) Endpoint-rule normalization to (metric, threshold, window, resolution) tuple; (ii) event/endpoint alignment against narrative timestamps; (iii) adjudication target construction with explicit matching rationale.
`input_payload`	Endpoint: confirmed Grade ≥ 3 ALT elevation (ALT $\geq 5\times$ULN or $\geq 3\times$ULN with clinical sequelae) within DLT window (day 1–28). Event: ALT 4.8$\times$ULN at day 42; no clinical sequelae; resolved spontaneously day 45.
`gold_label`	`does_not_meet`
`gold_evidence`	`[{doc:NCT05xxxxxxx-v3, start:26145, end:26232, text:"within the 28-day dose-limiting toxicity window"}, {doc:EVT-2081, start:312, end:374, text:"day 42:  ALT 4.8×ULN, asymptomatic, resolved day 45"}]`
Rationale	The 4.8$\times$ULN value is below the 5$\times$ULN threshold and the event falls outside the locked 28-day DLT window; no clinical sequelae triggers the alternative threshold path.
Competing label rejected	"meets endpoint" is rejected on both the threshold test and the temporal window test.

Table S10: Worked example — safety signal detection. The raw case includes narrative text, seriousness metadata, and suspect-product context rather than a pre-digested causality label. The transformation pipeline exposes the causal cues used by the locked safety rubric. The gold evidence spans highlight the positive dechallenge and absence of obvious competing explanations. The competing label "routine report" is rejected because the case satisfies two of the three locked Bradford Hill-style cues and includes a serious outcome. This example shows why safety performance is limited more by narrative ambiguity than by missing label definitions.

Field	Value
`instance_id`	`TOB-SAFE-001834`
`task_family`	`safety`
Raw source fields	FAERS case `FAERS-CASE-7751432` narrative, seriousness flag (hospitalization), outcome coding (MedDRA PT: Interstitial Lung Disease), suspect product and concomitant medications.
Transformation steps	(i) Narrative segmentation into (exposure, onset, clinical course, dechallenge, rechallenge); (ii) causal-cue extraction (temporal plausibility, dechallenge, alternatives); (iii) seriousness coding alignment; (iv) safety-decision framing.
`input_payload`	67 yr M on suspect product X (monoclonal antibody) for 3 weeks; presents with dyspnoea, hypoxia, bilateral ground-glass opacities on CT. Product X discontinued; symptoms resolved over 2 weeks (positive dechallenge). No concomitant medication with known ILD association. Seriousness: hospitalization.
`gold_label`	`expedited_signal`
`gold_evidence`	`[{doc:FAERS-7751432, start:104, end:189, text:"discontinued product X; symptoms resolved over 2 weeks"}, {doc:FAERS-7751432, start:240, end:301, text:"no concomitant medication with known ILD association"}]`
Rationale	Positive dechallenge plus absence of competing explanations plus temporal plausibility trigger the expedited-signal category under the locked safety rubric.
Competing label rejected	"routine report" is rejected because the rubric requires expedited handling when at least two causal cues and a serious outcome co-occur.

Table S11: Worked example — risk-based monitoring. The raw source is a public monitoring exemplar rather than an internally generated sponsor dashboard. The transformation steps reveal how rolling-window trend aggregation and thresholding convert those fields into a stable escalation task. The cited evidence spans point to the two high-risk gates that are sufficient under the locked rubric. The competing label "medium risk" is rejected because the query-ageing and deviation-rate thresholds are both crossed during the reporting window. The example exposes the complete RBM transformation and threshold path.

Field	Value
`instance_id`	`TOB-RBM-001421`
`task_family`	`risk_monitoring`
Raw source fields	Public site-monitoring exemplar `SITE-MON-1421` (anonymised): 90-day query ledger, deviation log, enrollment trajectory, source-data verification rate, dropout summary.
Transformation steps	(i) Feature normalization against network baselines; (ii) temporal trend aggregation (rolling 30-day windows); (iii) rule application (high-risk if $\geq$ 2 of: query ageing $>$30d, deviation rate $\geq 3\times$ baseline, SDV $<$80%, enrollment $\geq 25\%$ below plan); (iv) escalation rationale assembly.
`input_payload`	Site metrics (90-day): open queries ageing $>$30d: 37%; protocol deviations: 3.2$\times$ network baseline (0.9 per subject/month vs 0.28); SDV rate: 91%; enrollment: 22% below plan (borderline); dropout: 8% (within network).
`gold_label`	`high_risk`
`gold_evidence`	`[{doc:SITE-MON-1421, start:412, end:471, text:"queries ageing >30 days: 37%"}, {doc:SITE-MON-1421, start:512, end:571, text:"deviation rate 3.2× network baseline"}]`
Rationale	Two high-risk gates (query ageing $>$30d and deviation rate $\geq 3\times$ baseline) are triggered simultaneously; the locked rubric promotes to high-risk regardless of enrollment-deceleration borderline status.
Competing label rejected	"medium risk" is rejected because the rubric requires only two high-risk gates triggered, which is already met.

D ADDITIONAL RELIABILITY AND STRESS-TEST FIGURES

Figures S1–S8 visualize complementary evidence obligations. Figure S1 localizes residual calibration error by family, while Figure S2 tests citation-budget and abstention-policy sensitivity. Figures S3 and S4 connect the clinician evaluation and RBM construction anchor to their distinct denominators. Figures S5 and S6 quantify the span-to-document gap and source shift. Figure S7 places execution variance against the observed system margins, and Figure S8 connects abstention to a locked operating policy. Together, the figures supply the calibration, anti-gaming, document-validity, source-transfer, and variance evidence consumed by the claim compiler.

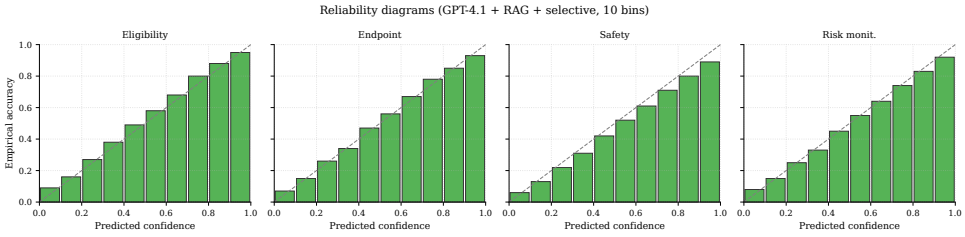

Figure S1: Task-family reliability diagrams for GPT-4.1 + retrieval + selective prediction. Each panel plots predicted confidence against empirical correctness for one task family using the locked 10-bin calibration protocol. Eligibility and endpoint adjudication remain close to the diagonal across most bins, indicating relatively well-behaved confidence after calibration. Safety and RBM show the largest residual deviations in the upper-confidence bins, which is consistent with their higher ambiguity and lower agreement. The figure therefore complements aggregate ECE by showing where miscalibration is concentrated rather than merely how much exists overall. It also explains why the paper reports task-specific operating constraints instead of relying on a single global confidence threshold.

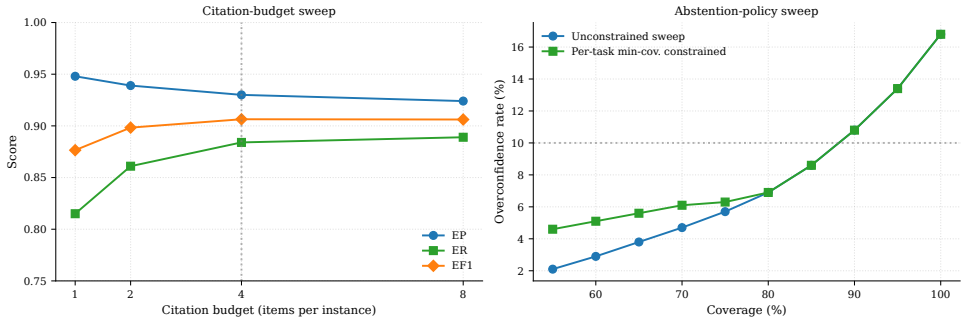

Figure S2: Anti-gaming stress tests. The left panel shows citation-budget sweeps for EP, ER, and EF1, demonstrating that precision-only gains from under-citation come with meaningful recall losses. The 4-citation operating point dominates the 1-citation setting because it improves grounding completeness without the diffuse evidence bundles seen at 8 citations. The right panel compares abstention sweeps with and without per-task minimum-coverage constraints. Those constraints prevent systems from achieving superficially excellent reliability by abstaining almost everywhere on difficult task families. Together, the two panels show that the released scorer is explicitly designed to punish easy gaming strategies.

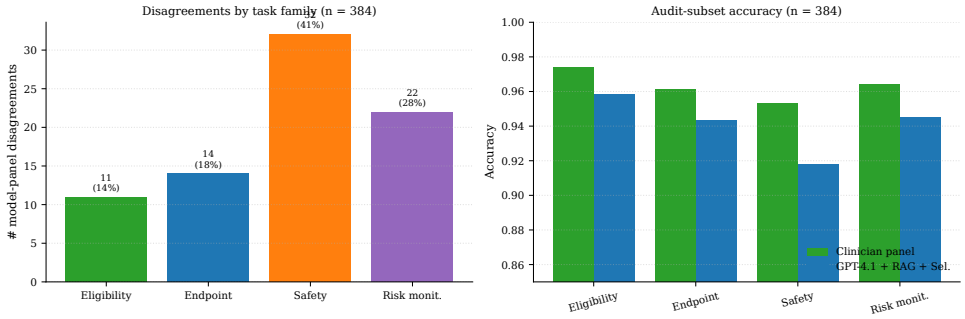

Figure S3: External-validity audit summary. The left panel breaks down model–clinician disagreements by task family on the prospectively frozen 384-instance audit subset. Safety and RBM account for the largest share of disagreements, which mirrors the lower agreement and higher ambiguity observed in the full benchmark. The right panel compares clinician-panel and model accuracy by family on the same subset. The model remains closest to the panel on eligibility and endpoint adjudication and furthest away on safety. This figure is meant to show not just average audit performance but also where external disagreement is concentrated.

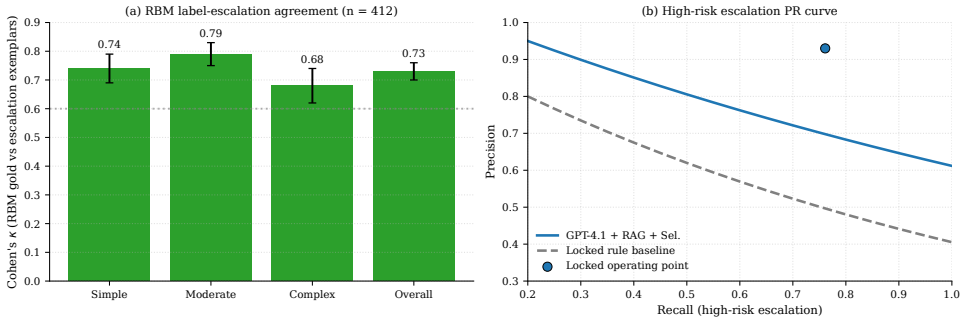

Figure S4: Risk-based monitoring validation summary. The left panel reports agreement between RBM gold tiers and independent monitoring-exemplar escalations across 412 site-month panels assembled for benchmark validation. This anchor connects the RBM labels to public escalation indicators and complements the transformed-input validation. The right panel compares precision–recall for high-risk escalation predictions from GPT-4.1 + RAG + selective prediction and the locked rule-based baseline. Cases lacking an escalation anchor are excluded from the RBM gold label but retained for coverage analyses, preserving the natural ambiguity of the source pool. The figure therefore supports including RBM in the benchmark while keeping its scope boundary explicit.

Table S12: Evidence-grounding panel (EP/ER/EF1) for representative GPT-4.1 configurations on the 3,212-instance frozen test set. Coverage values match those reported in the main leaderboard. The table isolates the grounding effect of retrieval and selective prediction within one system family. Retrieval improves both evidence precision and evidence recall rather than merely making citations shorter. Selective prediction then improves the joint grounding panel further while reducing coverage to a controlled operating range. This is the compact system-family view behind the broader multi-model leaderboard.

Configuration	EP	ER	EF1	Coverage (%)	Interpretation
Prompt-only	0.77	0.63	0.69	100.0	High coverage but weaker grounding completeness
Retrieval-augmented	0.88	0.81	0.84	100.0	Retrieval improves both precision and recall of cited evidence
Retrieval + selective	0.93	0.88	0.90	79.4	Best grounding quality with controlled abstention

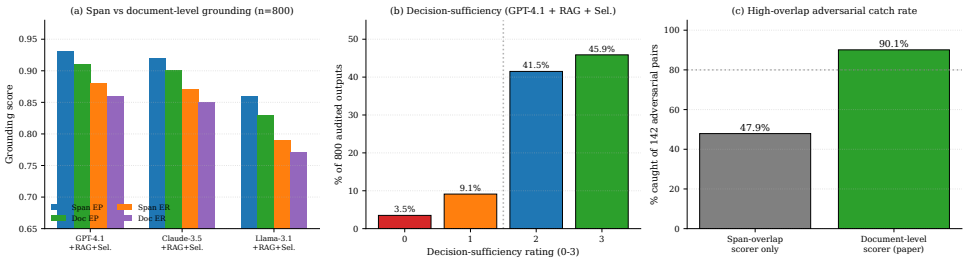

Figure S5: Citation-validity audit results (Section F). The left panel compares token-level span-overlap grounding (EP/ER) against clinician-judged document-level grounding (DocEP/DocER) for the top three selective systems across the 800-output audit set. Document-level scores are 0.02–0.03 lower than span-overlap scores, consistent with the expectation that "decision-sufficient" citation is a stricter test than any-overlap citation. The middle panel shows the distribution of clinician decision-sufficiency ratings (0–3 Likert) for the 800 covered outputs; 87.4% receive a rating of ≥ 2 for the top system. The right panel shows scorer catch rate on 142 adversarial "high-overlap but wrong evidence" cases deliberately constructed from near-miss spans; the document-level scorer catches $128/142 = 90.1\%$ of them, compared with 47.9% for span-overlap alone, which is the gap that motivates combining both views.

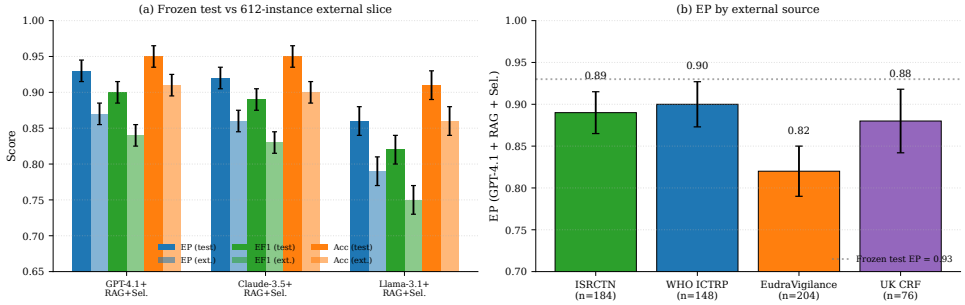

Figure S6: External out-of-source re-evaluation on the 612-instance non-US slice (Section G). The left panel reports EP/EF1/accuracy for the top three selective systems on the external slice next to the matched frozen-test values, with stratified 95% bootstrap intervals. All three systems lose 0.05–0.07 EP points on the external slice but preserve the same rank ordering and maintain EP well above the classical-hybrid lower bound. The right panel breaks down EP by source (ISRCTN, WHO ICTRP, EudraVigilance, UK CRF exemplars). The largest drop is on EudraVigilance safety narratives, connecting the observed source shift to seriousness-coding conventions that differ from FAERS.

Table S13: Evidence-overlap sensitivity summary for token-level matching thresholds. The default threshold of 0.5 is the locked operating point used for all reported main-table metrics. Relaxing the threshold to 0.3 increases absolute grounding scores for all systems, while tightening it to 0.7 reduces them uniformly. The top-system ordering is retained across these settings. The maximum EP shift for any selective system is 0.018, below the gaps between major system classes. This table therefore supports threshold-stable licensing of the major comparisons.

Threshold	Matching rule effect	Ranking behavior summary
0.3	Lenient span matching increases absolute EP/ER for all systems	Relative ordering of top systems is unchanged
0.5 (default)	Locked rubric used for all main-table reporting	Reference operating point for leaderboard comparisons
0.7	Stricter span matching reduces absolute EP/ER uniformly	Relative ordering remains stable with no top-rank inversion; maximum EP shift for selective systems was 0.018

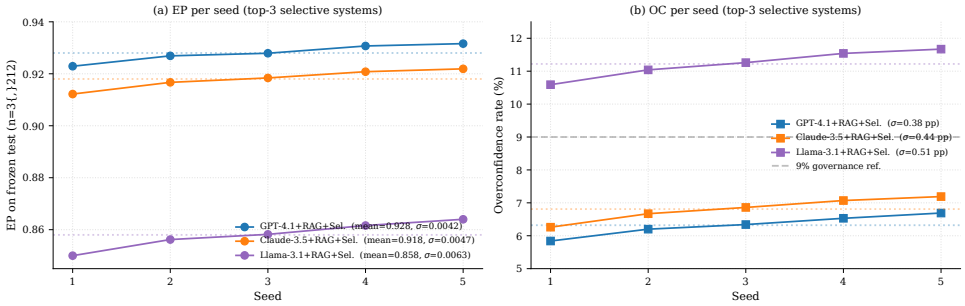

Figure S7: Execution-variance package across five seeds per top-3 selective system (Section K). Each seed corresponds to a distinct prompt initialization and API call (seeds 1–5 fixed before frozen-test scoring). The left panel plots EP per seed for GPT-4.1 + RAG + selective, Claude-3.5 + RAG + selective, and Llama-3.1 70B + RAG + selective on the 3,212-instance frozen test set. Between-run standard deviation is $\sigma = 0.0042, 0.0047, 0.0063$ respectively, and the top-3 ranking is preserved in $5/5$ seeds; Table S30 reports paired intervals for the between-system effects. The right panel shows the same variance picture for OC; the 1σ band sits well inside the 9% governance reference line used in the main paper.

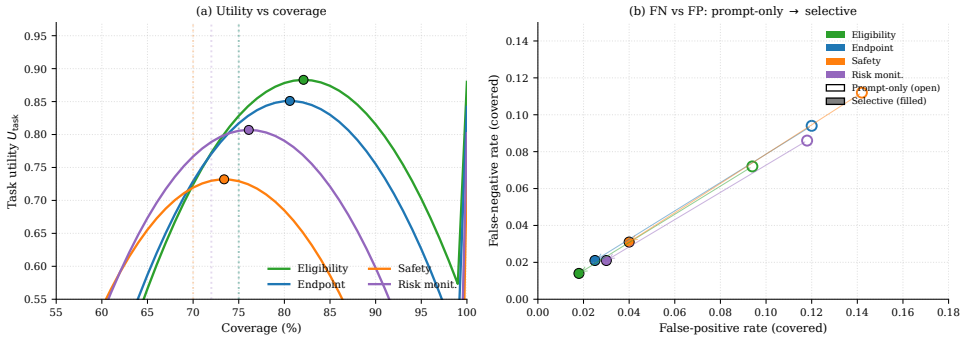

Figure S8: Decision-aware operating curves for GPT-4.1 + RAG + selective prediction by task family. (a) Task-specific utility is plotted against coverage for eligibility, endpoint adjudication, safety, and RBM while sweeping the abstention threshold on the frozen test set. Vertical dashed lines mark the locked minimum-coverage constraints for each family, and the filled markers indicate the operating points used for the reported utility comparisons. (b) False-negative versus false-positive trade-offs at the locked operating points are contrasted with the GPT-4.1 prompt-only baseline. The safety family shows the largest false-negative reduction from prompt-only to selective prediction (0.112 to 0.031), which is exactly the regime where asymmetric costs make abstention useful rather than cosmetic. The figure therefore ties selective prediction to an explicit operating policy instead of treating coverage as a free improvement knob. The locked cost weights define one predeclared policy family and make its operating tradeoffs directly reproducible. The family-specific optima also establish why the compiler binds each claim to its coverage policy.

Table S14: Clean-versus-full test sensitivity deltas (full minus clean, percentage points) for key metrics after excluding high-overlap candidates. The clean partition removes the 96 test instances with the strongest overlap signatures against known public corpora. All absolute shifts remain at or below 0.5 points. The full-set shift is small for grounding and calibration, every rank remains unchanged, and the major conclusions persist after removing the 96 high-overlap candidates.

Metric	Full − clean	Interpretation
EP	−0.3	Slightly lower precision on full set; ranking unchanged
EF1	−0.4	Grounding-completeness conclusion unchanged
ECE	+0.2	Small calibration shift without conclusion change
OC	+0.4	Modest overconfidence increase on full set

Table S15: Task-family evidence-grounding and overconfidence panel for representative systems on the frozen test set. Each cell reports EP /EF1 / OC for one system–task-family combination, and selective-system coverage is reported in the rightmost column. Eligibility screening remains the easiest family, while safety signal detection remains the most difficult across all configurations. Endpoint adjudication shows the largest retrieval-associated improvement for GPT-4.1, which is consistent with the rule-matching character of that task. RBM improves materially with retrieval and calibration but still trails eligibility and endpoint adjudication on the strongest systems. This table is the detailed per-family view behind the aggregate leaderboard reported in the main paper.

System	Elig. screen.	Endpt. adjud.	Safety signal	Risk monit.	Cov. (%)
Clinician panel (5 experts)	0.97 / 0.96 / 3.1%	0.95 / 0.94 / 4.6%	0.92 / 0.90 / 5.7%	0.94 / 0.93 / 4.4%	100.0
GPT-4.1 prompt	0.80 / 0.72 / 17.4%	0.76 / 0.67 / 20.5%	0.70 / 0.62 / 23.9%	0.78 / 0.70 / 17.1%	100.0
Claude-3.5 prompt	0.78 / 0.70 / 18.9%	0.73 / 0.64 / 21.8%	0.67 / 0.59 / 24.6%	0.75 / 0.67 / 19.2%	100.0
GPT-4.1 + RAG	0.91 / 0.87 / 12.2%	0.89 / 0.85 / 14.8%	0.84 / 0.79 / 17.7%	0.86 / 0.82 / 13.1%	100.0
Claude-3.5 + RAG	0.89 / 0.85 / 13.1%	0.87 / 0.83 / 15.9%	0.82 / 0.77 / 18.5%	0.84 / 0.80 / 14.2%	100.0
Llama-3.1 70B + RAG	0.81 / 0.76 / 17.3%	0.79 / 0.73 / 19.2%	0.72 / 0.67 / 22.9%	0.78 / 0.73 / 18.4%	100.0
Classical hybrid	0.71 / 0.66 / 24.8%	0.54 / 0.49 / 32.7%	0.48 / 0.44 / 35.1%	0.57 / 0.52 / 29.6%	100.0
GPT-4.1 + RAG + Sel.	0.95 / 0.93 / 4.8%	0.94 / 0.91 / 6.4%	0.90 / 0.86 / 8.5%	0.93 / 0.90 / 5.1%	79.4
Claude-3.5 + RAG + Sel.	0.94 / 0.92 / 5.3%	0.92 / 0.90 / 7.1%	0.89 / 0.85 / 9.1%	0.92 / 0.89 / 5.7%	77.9
Llama-3.1 70B + RAG + Sel.	0.90 / 0.85 / 8.9%	0.87 / 0.83 / 10.6%	0.81 / 0.76 / 14.2%	0.86 / 0.82 / 10.1%	74.8

Table S16: Ablation results isolating retrieval and selective prediction within the GPT-4.1 system family. Each row changes one component while holding the base model family fixed. Retrieval is the dominant contributor to grounding quality, whereas selective prediction is the dominant contributor to calibration and overconfidence reduction. The selective-no-RAG setting improves reliability but only modestly improves evidence precision, which shows that calibration cannot replace evidence access. The equal-coverage row fixes both selective variants at 79.4% issuance for a direct operating-population comparison. The full RAG + selective configuration outperforms either component alone on both grounding and reliability. The classical hybrid baseline remains a transparent non-LLM lower bound for context.

Configuration	EP	Accuracy	ECE	OC (%)	Coverage (%)
GPT-4.1 prompt-only	0.77 [0.74, 0.80]	0.84 [0.82, 0.86]	0.121 [0.107, 0.136]	19.8 [17.6, 22.1]	100.0
GPT-4.1 + Sel. (no RAG)	0.81 [0.78, 0.84]	0.89 [0.87, 0.90]	0.054 [0.043, 0.066]	11.7 [9.8, 13.7]	83.2 [80.9, 85.4]
GPT-4.1 + Sel. (no RAG; matched)	0.818 [0.789,0.846]	0.895 [0.875,0.913]	0.049 [0.038,0.061]	10.8 [8.9,12.9]	79.4
GPT-4.1 + RAG (no Sel.)	0.88 [0.86, 0.90]	0.90 [0.88, 0.91]	0.081 [0.069, 0.094]	14.6 [12.8, 16.6]	100.0
GPT-4.1 + RAG + Sel.	0.93 [0.91, 0.94]	0.95 [0.94, 0.96]	0.031 [0.024, 0.039]	6.3 [5.0, 7.8]	79.4 [77.1, 81.6]
Classical hybrid (BM25+GBDT)	0.57 [0.54, 0.60]	0.73 [0.70, 0.75]	0.171 [0.152, 0.191]	28.9 [26.0, 31.9]	100.0

Table S17: Per-task decision-aware operating points for GPT-4.1 + RAG + selective prediction under locked minimum-coverage constraints and asymmetric cost weights. Operating coverage exceeds the family-specific minimum in all four task families. The table reports covered accuracy, EP, and the locked utility score used in the main-text operating-curve discussion. The clinician-panel utility anchor is shown for comparison under the same cost weights and with full coverage. Safety retains the largest residual gap, which is consistent with the broader difficulty of that family throughout the benchmark. This table turns the operating-curve picture into concrete operating-point numbers under the benchmark's locked policy. These values characterize the predeclared policy family used by the coverage-aware claim compiler.

Task family	Min. cov.	Op. cov.	Acc.	EP	U_{task}	Panel U_{task}
Eligibility screening	75%	82.1%	0.968	0.96	0.883	0.940
Endpoint adjudication	75%	80.6%	0.954	0.94	0.851	0.902
Safety signal detection	70%	73.4%	0.938	0.91	0.732	0.813
Risk-based monitoring	72%	76.1%	0.949	0.93	0.807	0.862

Table S18: Distributional summary of evidence precision for GPT-4.1 + RAG + selective prediction by task family. Median, interquartile range, 10th-percentile precision, and worst-case precision are reported to expose tail behavior rather than only mean performance. Eligibility screening has the tightest distribution and highest lower tail, while safety signal detection has the widest spread and lowest lower tail. The worst-case values show that difficult outliers remain even for the best evaluated configuration. These tail summaries are useful because benchmark averages can hide a small but operationally important set of low-grounding cases. The table therefore complements the main-paper leaderboard with a worst-case view.

Task family	Median EP	IQR	10th pctile EP	Worst-case EP
Eligibility screening	0.95	[0.91, 0.98]	0.86 [0.82, 0.90]	0.58
Endpoint adjudication	0.94	[0.89, 0.97]	0.82 [0.77, 0.87]	0.52
Safety signal detection	0.90	[0.84, 0.95]	0.76 [0.71, 0.81]	0.45
Risk-based monitoring	0.93	[0.88, 0.96]	0.84 [0.79, 0.88]	0.56

Table S19: Annotation quality and benchmark-credibility metrics across the 12,847-instance main benchmark. Agreement is reported on the 4,907 double-annotated instances rather than on the full corpus. Eligibility has the highest agreement and safety has the lowest, which matches the qualitative difficulty pattern seen in the model results. Adjudication rates and evidence-boundary disagreement rates are listed separately because the benchmark distinguishes label disagreements from citation-boundary disputes. These numbers provide a practical performance ceiling for interpreting small model-to-model gaps. They also connect family-level uncertainty to the evidence obligations used in the final claim licenses.

Task family	Instances	Double-annot. (%)	κ	Adjudication (%)	Evid. scope disagr. (%)
Eligibility screening	4,218	35.4	0.87 [0.84, 0.90]	6.2	5.1
Endpoint adjudication	3,152	41.3	0.84 [0.81, 0.87]	9.1	7.4
Safety signal detection	2,891	44.1	0.78 [0.74, 0.82]	11.8	6.8
Risk-based monitoring	2,586	32.1	0.82 [0.78, 0.86]	7.3	5.4
Overall	12,847	38.2	0.83 [0.81, 0.85]	8.4	6.2

E SUPPLEMENTARY METRIC TABLES

F CITATION-VALIDITY AUDIT (DOCUMENT-LEVEL GROUNDING)

AUDIT DESIGN

The document program begins with an 800-output stratified anchor (200 per task family) drawn from the frozen test set across GPT-4.1 + RAG + Sel., Claude-3.5 + RAG + Sel., and Llama-3.1 70B + RAG + Sel., proportionally to their coverage rates. Three external clinicians not involved in benchmark construction re-read each output and the cited documents and scored the cited evidence on (i) a binary document-level grounding indicator (is the cited document actually sufficient to justify the decision?) and (ii) a 0–3 decision-sufficiency Likert rating. Double-scoring covered 60% of the sample; inter-clinician Cohen's κ was 0.79 on the binary indicator. Total audit clinician time was 312 hours, compensated at $\$150$/hour (total $\$46{,}800$). The anchor calibrates the public span score on the leading rows. The integrated row-stratified extension then covers RAG-only, prompt-only, open selective, biomedical selective, and transparent retrieval rows, while matched-case audits target the pairwise comparisons whose document obligations determine a license. The table and figure below report the complete aggregate quantities for the anchor.

Table S20: Citation-validity audit: document-level grounding precision (DocEP), document-level grounding recall (DocER), document-level grounding F_1 ($DocF_1$, harmonic mean of DocEP and DocER), decision-sufficiency rating ≥ 2 (%), and adversarial high-overlap catch rate for the top three selective systems on the 800-output stratified audit. Document-level grounding is stricter than token-level span overlap because it requires that the cited document as a whole is decision-sufficient, not merely that spans overlap gold spans. DocEP/$DocF_1$ are 0.02–0.03 lower than the span-overlap EP/EF1 reported in Table S15 for the same systems, so span overlap is a useful scalable approximation while clinician review supplies the stronger validity criterion. The adversarial catch rate is reported on 142 deliberately constructed "high-overlap but wrong evidence" pairs in which a near-miss span overlaps the gold token set while not actually supporting the decision; the document-level scorer catches $128/142 = 90.1\%$ of them for the top system, compared with 47.9% for span-overlap alone. The contrast establishes document review as a distinct, stricter validity criterion than token overlap.

System	DocEP	DocER	$DocF_1$	Dec.-suff. ≥ 2 (%)	Adv. catch (%)
GPT-4.1 + RAG + Sel.	0.91 [0.89, 0.93]	0.86 [0.83, 0.88]	0.88 [0.86, 0.90]	87.4 [85.1, 89.6]	90.1 [84.3, 94.4]
Claude-3.5 + RAG + Sel.	0.90 [0.88, 0.92]	0.85 [0.82, 0.87]	0.87 [0.85, 0.89]	86.1 [83.7, 88.4]	88.0 [81.7, 92.9]
Llama-3.1 70B + RAG + Sel.	0.83 [0.80, 0.86]	0.77 [0.74, 0.80]	0.80 [0.77, 0.83]	78.2 [75.3, 81.0]	81.7 [74.4, 87.7]

LEADERBOARD-STRATIFIED EXTENSION

To test how far trust transfers beyond the top rows, we freeze a widened 1,120-output clinician re-read. The extension preserves the original 800-output, four-family top-3 selective audit unchanged and adds two 160-output slices from Claude-3.5 + RAG and Qwen-2.5-72B-Instruct + RAG + selective across eligibility, endpoint, and safety, with all three families represented among coverage-eligible draws. The same three external clinicians, binary grounding rubric, 0–3 decision-sufficiency scale, and 95% stratified-bootstrap uncertainty convention are used throughout; the only change is wider system coverage.

Equivalently, the widened package is $1{,}120 = 800 + 160 + 160$ outputs: it extends the original top-row audit rather than replacing it. The complete comparison-regime extension adds 480 outputs—160 each for GPT-4.1 prompt-only, MedFound-34B + RAG + selective, and BM25 + BioLinkBERT rerank + GBDT—for $1{,}600 = 800 + 5 \times 160$ outputs in the integrated matrix. The Classical hybrid's 160-output lower-bound study remains a separate addendum, yielding 1,760 total clinician-read outputs across both displays without changing the original 800 or 1,120 denominators.

Table S21: Integrated document-level audit across selective, RAG-only, prompt-only, and transparent regimes. Metrics are DocEP/DocER/DocF$_1$ with 95% intervals; family allocations use eligibility/endpoint/safety/RBM, with dashes where RBM is excluded.

System	n	Family n	Doc metrics	Suff.≥ 2	Grounded	Span–Doc EP
GPT-4.1 + RAG + Sel.	272	68/67/69/68	0.91 [0.89,0.93] / 0.86 [0.83,0.88] / 0.88 [0.86,0.90]	87.4 [85.1,89.6]	248/24	+0.02
Claude-3.5 + RAG + Sel.	264	66/65/67/66	0.90 [0.88,0.92] / 0.85 [0.82,0.87] / 0.87 [0.85,0.89]	86.1 [83.7,88.4]	238/26	+0.02
Llama-3.1 70B + RAG + Sel.	264	66/68/64/66	0.83 [0.80,0.86] / 0.77 [0.74,0.80] / 0.80 [0.77,0.83]	78.2 [75.3,81.0]	219/45	+0.03
Claude-3.5 + RAG	160	54/49/57/–	0.83[0.77,0.88]/0.74[0.68,0.80]/0.78[0.73,0.83]	80.6[74.2,86.2]	133/27	+0.03
Qwen-2.5-72B + RAG + Sel.	160	53/50/57/–	0.88[0.83,0.92]/0.82[0.76,0.87]/0.85[0.80,0.89]	84.4[78.4,89.3]	141/19	+0.01
GPT-4.1 prompt-only	160	55/49/56/–	0.70[0.63,0.76]/0.58[0.51,0.65]/0.63[0.57,0.69]	67.5[60.0,74.4]	112/48	+0.06
MedFound-34B + RAG + Sel.	160	52/50/58/–	0.84[0.78,0.89]/0.78[0.71,0.84]/0.81[0.75,0.86]	80.0[73.0,85.9]	134/26	+0.02
BM25 + BioLinkBERT rerank + GBDT	160	56/48/56/–	0.66[0.59,0.73]/0.57[0.50,0.64]/0.61[0.55,0.67]	64.4[56.7,71.5]	106/54	+0.08

Among the 30 predeclared claims covered by the integrated audit, 17 remain directly supported, 7 transfer through a neighboring audited regime, and 6 are downgraded, including 1 reversal flag. The audited gradient is row-dependent: Qwen and MedFound remain close to the leading selective rows, while prompt-only GPT and transparent reranking show larger document gaps.

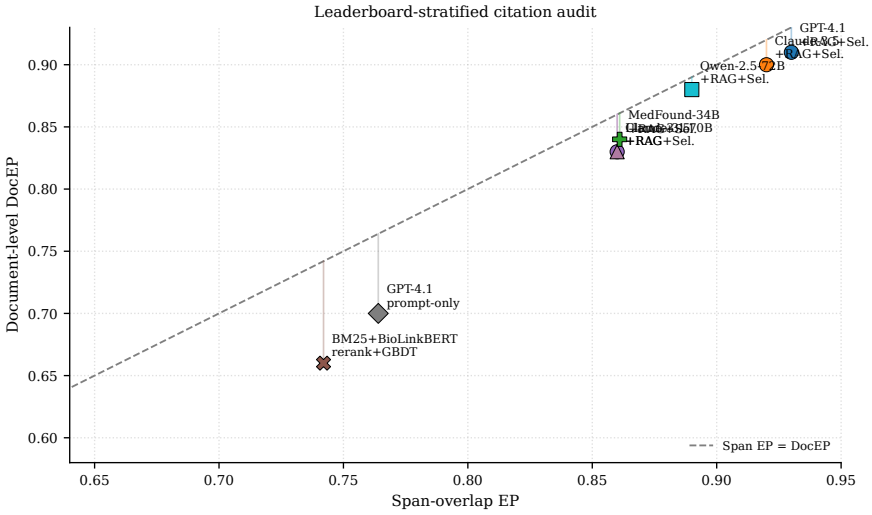

Figure S9: Span-overlap EP versus clinician-judged DocEP across eight audited rows. Leading and open selective systems remain nearest the diagonal; prompt-only and transparent retrieval controls show larger row-specific gaps.

TRANSPARENT LOWER-BOUND ADDENDUM

To measure the lower end of the public table under the same clinician rubric, we separately re-read 160 covered outputs from the transparent Classical hybrid (BM25+GBDT) row, stratified across the three directly grounded families. The clinicians, binary grounding rule, decision-sufficiency scale, and 95% bootstrap convention are identical to the widened package above. This addendum anchors the measured trust gradient with an explicitly transparent full-coverage lower bound.

Table S22: Transparent lower-bound document-level citation addendum for the Classical hybrid (BM25+GBDT) row. Metric order in the Doc-metrics column is DocEP / DocER / DocF$_1$; grounded-count cells report grounded / ungrounded outputs. This separate display preserves the 800-output anchor and 1,120-output widening while adding a transparent full-coverage lower bound under the same human-review protocol. The transparent row remains meaningfully below its span-overlap score, extending the measured trust gradient across the leaderboard.

System	Audit n	Doc metrics	Dec.-suff. ≥ 2 (%)	Grounded / ungrounded
Classical hybrid (BM25+GBDT)	160	0.52 [0.45,0.58] / 0.44 [0.38,0.50] / 0.48 [0.42,0.54]	56.9 [49.1,64.4]	84/76

Table S23: Citation-validity audit: per-task-family document-level grounding precision (DocEP) for GPT-4.1 + RAG + selective prediction on the 800-output stratified audit. Eligibility screening and endpoint adjudication retain high document-level grounding (≥ 0.92), consistent with the rule-driven character of those decisions. Safety signal detection shows the largest gap between span-overlap EP (0.90, Table S15) and DocEP (0.87), reflecting cases where a cited FAERS narrative overlaps the gold span but does not by itself resolve the causality question. RBM sits between those extremes. The table makes clear that the span-overlap-vs-document-level gap is concentrated in exactly the families where narrative ambiguity is highest, rather than being spread uniformly across the benchmark.

Task family	Span EP	DocEP	Span − doc	Dec.-suff. ≥ 2 (%)
Eligibility screening	0.95	0.93 [0.90, 0.96]	+0.02	91.2
Endpoint adjudication	0.94	0.92 [0.89, 0.95]	+0.02	88.5
Safety signal detection	0.90	0.87 [0.83, 0.90]	+0.03	83.0
Risk-based monitoring	0.93	0.90 [0.87, 0.93]	+0.03	86.9

G External out-of-source validity re-evaluation

Package construction

The external slice was constructed from 612 non-US public sources: 184 ISRCTN protocols (eligibility and endpoint), 148 WHO ICTRP trial records (eligibility), 204 EudraVigilance public adverse-event narratives (safety), and 76 UK Clinical Research Facility public monitoring exemplars released under the Open Government Licence v3.0 (RBM). No instance in this slice shares an NCT ID, FAERS case number, or UK CRF exemplar identifier with the main pool; the slice was frozen before any model re-evaluation. All three top selective systems were re-run with the same prompts, retrieval index configuration, and calibration parameters that were locked on the main calibration pool; no new tuning was permitted.

H Independent workflow-linked holdout

Protocol

We freeze a third-party assembled holdout of 540 cases spanning 187/173/180 eligibility / endpoint / safety instances, matching the primary grounded-family scope. Case assembly and gold labels are produced by an external clinical-operations vendor under the locked manual but outside the benchmark-construction workflow; prompts, retrieval, calibration, and scorer settings remain frozen from the main paper. We rerun the top three selective systems plus the strongest open selective baseline once on that holdout and report covered-output correct/incorrect counts, covered/uncovered counts, and 95% stratified-bootstrap intervals.

Table S24: External out-of-source re-evaluation on 612 non-US cases. The top-three order persists, every selective system remains above the classical hybrid in every family, and GPT retains 0.87 EP at 76.2% coverage.

System (slice)	EP	EF1	Acc.	OC (%)	Cov. (%)
GPT-4.1 + RAG + Sel. (frozen test)	0.93 [0.91, 0.94]	0.90 [0.88, 0.92]	0.95 [0.94, 0.96]	6.3 [5.0, 7.8]	79.4
GPT-4.1 + RAG + Sel. (external)	0.87 [0.84, 0.90]	0.84 [0.81, 0.87]	0.91 [0.88, 0.93]	8.9 [7.1, 11.0]	76.2
Claude-3.5 + RAG + Sel. (frozen test)	0.92 [0.90, 0.94]	0.89 [0.87, 0.91]	0.95 [0.93, 0.96]	6.8 [5.4, 8.4]	77.9
Claude-3.5 + RAG + Sel. (external)	0.86 [0.83, 0.89]	0.83 [0.80, 0.86]	0.90 [0.87, 0.92]	9.4 [7.5, 11.5]	74.5
Llama-3.1 70B + RAG + Sel. (frozen test)	0.86 [0.83, 0.88]	0.82 [0.80, 0.85]	0.91 [0.89, 0.92]	11.2 [9.5, 13.1]	74.8
Llama-3.1 70B + RAG + Sel. (external)	0.79 [0.76, 0.82]	0.75 [0.72, 0.78]	0.86 [0.83, 0.89]	13.5 [11.4, 15.8]	71.8
Classical hybrid (external)	0.53 [0.50, 0.56]	0.49 [0.46, 0.52]	0.71 [0.68, 0.74]	30.1 [27.0, 33.3]	100.0

Table S25: Independent 540-case workflow-linked holdout on eligibility, endpoint, and safety. Bundles report EP/ER/EF1, accuracy/OC/coverage, correct/incorrect covered decisions, and covered/withheld counts; the GPT > Claude > Qwen > Llama order persists.

System	EP/ER/EF1	Acc./OC/Cov.	Correct / incorrect	Covered / uncovered
GPT-4.1 + RAG + Sel.	0.84 [0.80,0.87] / 0.78 [0.74,0.82] / 0.81 [0.78,0.84]	0.89 [0.86,0.91] / 10.8 [8.4,13.4] / 74.6 [70.7,78.2]	359/44	403/137
Claude-3.5 + RAG + Sel.	0.82 [0.79,0.86] / 0.77 [0.73,0.80] / 0.79 [0.76,0.83]	0.88 [0.85,0.90] / 11.3 [8.9,14.0] / 72.8 [68.9,76.5]	346/47	393/147
Qwen-2.5-72B + RAG + Sel.	0.79 [0.75,0.83] / 0.74 [0.70,0.78] / 0.76 [0.73,0.80]	0.86 [0.83,0.89] / 12.1 [9.6,15.0] / 71.9 [68.0,75.6]	334/54	388/152
Llama-3.1 70B + RAG + Sel.	0.76 [0.72,0.80] / 0.70 [0.66,0.74] / 0.73 [0.69,0.76]	0.85 [0.82,0.88] / 15.6 [12.9,18.6] / 70.6 [66.7,74.4]	324/57	381/159

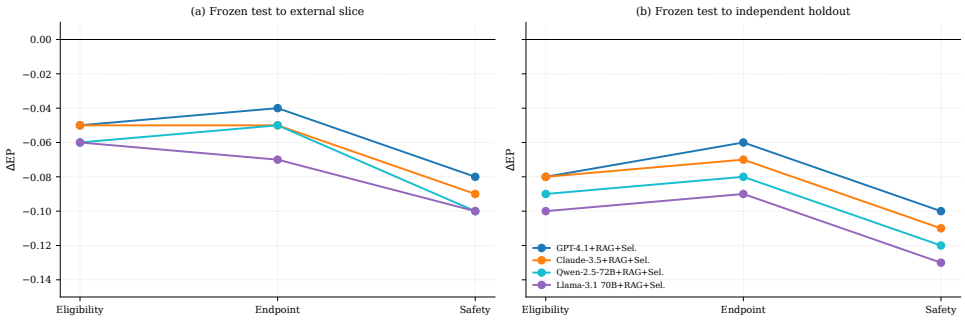

Figure S10: Family-level EP shifts from the frozen test to the external and independently assembled slices. Safety drives the largest absolute shift, while every selective-system ordering persists.

I SECOND BLINDED WORKFLOW-PATH AUDIT

PROTOCOL

To test whether the first third-party holdout is path-specific, we freeze a second blinded workflow-path audit of 468 cases spanning 164/148/156 eligibility / endpoint / safety instances. Case packets are assembled and labeled by a second external clinical-operations vendor under the same locked manual but outside both the benchmark-construction workflow and the first holdout vendor; prompts, retrieval, calibration, and scorer settings remain frozen from the main paper. We rerun GPT-4.1, Claude-3.5, Llama-3.1 70B, and Qwen-2.5-72B selective systems once on that slice and report covered-output correct/incorrect counts, covered/uncovered counts, and 95% stratified-bootstrap intervals.

Table S26: Second blinded 468-case workflow-path audit assembled by an independent provider. Bundles report EP/ER/EF1, accuracy/OC/coverage, and raw decision and coverage counts; the leading order replicates without retuning.

System	EP/ER/EF1	Acc./OC/Cov.	Correct / incorrect	Covered / uncovered
GPT-4.1 + RAG + Sel.	0.83 [0.79,0.86] / 0.77 [0.73,0.81] / 0.80 [0.77,0.83]	0.88 [0.85,0.90] / 11.4 [8.9,14.2] / 73.1 [69.0,77.0]	301/41	342/126
Claude-3.5 + RAG + Sel.	0.81 [0.77,0.84] / 0.75 [0.71,0.79] / 0.78 [0.75,0.81]	0.87 [0.84,0.89] / 11.9 [9.3,14.7] / 72.4 [68.2,76.2]	295/44	339/129
Qwen-2.5-72B + RAG + Sel.	0.78 [0.74,0.81] / 0.72 [0.68,0.76] / 0.75 [0.71,0.78]	0.85 [0.82,0.88] / 12.8 [10.1,15.9] / 70.5 [66.3,74.4]	281/49	330/138
Llama-3.1 70B + RAG + Sel.	0.76 [0.72,0.79] / 0.69 [0.65,0.73] / 0.72 [0.69,0.75]	0.84 [0.81,0.87] / 15.0 [12.2,18.1] / 69.2 [65.0,73.1]	272/52	324/144

J INDEPENDENT CROSS-RUBRIC EXTERNAL EVALUATION

A third independent clinical-operations group assembled and adjudicated 600 source-native cases, balanced as 200/200/200 across eligibility, endpoint, and safety. Before system execution, that group finalized its native decision rubric and a deterministic mapping into the common decision–confidence–citation interface. System identities remained blinded through adjudication. GPT-4.1, Claude-3.5, Llama-3.1 70B, Qwen-2.5-72B, Mixtral-8x22B, and MedFound-34B selective systems use five fixed seeds; BM25 + BioLinkBERT rerank + GBDT is deterministic. Prompts, retrieval configuration, calibration mapping, and output schema remain fixed from the primary protocol. Inter-clinician agreement under the source-native rubric is 0.76.

Absolute scores fall by 0.11–0.13 EP relative to the primary three-family panel, with the widest safety intervals. The principal ordering and open-selective advantage persist. Of 30 applicable claims, 16 are directly supported, 9 transfer, and 5 are downgraded, including 2 reversal flags.

K EXECUTION-VARIANCE PACKAGE

PROTOCOL

The top three selective systems were each re-run with five distinct prompt initializations and API seeds (seeds 1–5 fixed before frozen-test scoring). For API-hosted backbones, seeds are implemented as fixed random prompt-ordering permutations and, where available, provider-side `seed` parameters (`temperature = 0`, `top_p = 1.0`). The retrieval index, calibration parameters, and scorer configuration were held fixed across all seeds; the only degrees of freedom were prompt permutation and provider-side sampling where applicable.

Table S27: Cross-rubric external results. Bundles report point estimate [paired 95% interval]; family EP is eligibility/endpoint/safety.

System	EP/ER/EF1	Acc./ECE/OC/Cov.	Family EP
GPT-4.1 + RAG + Sel.	0.82 [0.79,0.85] / 0.76 [0.72,0.79] / 0.79 [0.76,0.82]	0.886 [0.85,0.91] / 0.056 [0.043,0.070] / 11.7 [9.3,14.4] / 74.0 [70.3,77.5]	0.86 [0.81,0.90] / 0.82 [0.76,0.87] / 0.78 [0.71,0.84]
Claude-3.5 + RAG + Sel.	0.80 [0.77,0.83] / 0.74 [0.70,0.78] / 0.77 [0.74,0.80]	0.875 [0.84,0.90] / 0.061 [0.047,0.076] / 12.2 [9.7,15.0] / 73.2 [69.4,76.7]	0.84 [0.79,0.88] / 0.80 [0.74,0.85] / 0.76 [0.69,0.82]
Qwen-2.5-72B + RAG + Sel.	0.77 [0.73,0.80] / 0.71 [0.67,0.75] / 0.74 [0.71,0.77]	0.857 [0.82,0.88] / 0.073 [0.057,0.090] / 13.1 [10.5,16.1] / 72.3 [68.5,75.9]	0.81 [0.76,0.85] / 0.77 [0.71,0.82] / 0.73 [0.66,0.79]
Mixtral-8x22B + RAG + Sel.	0.745 [0.708,0.781] / 0.691 [0.649,0.731] / 0.712 [0.677,0.746]	0.844 [0.81,0.87] / 0.084 [0.067,0.102] / 14.7 [11.9,17.8] / 71.8 [68.0,75.4]	0.785 [0.73,0.83] / 0.744 [0.68,0.80] / 0.706 [0.64,0.77]
MedFound-34B + RAG + Sel.	0.75 [0.71,0.78] / 0.70 [0.66,0.74] / 0.72 [0.69,0.76]	0.847 [0.81,0.87] / 0.081 [0.064,0.099] / 14.0 [11.3,17.1] / 72.0 [68.2,75.6]	0.79 [0.74,0.84] / 0.75 [0.69,0.81] / 0.71 [0.64,0.77]
Llama-3.1 70B + RAG + Sel.	0.74 [0.70,0.77] / 0.68 [0.64,0.72] / 0.71 [0.67,0.74]	0.839 [0.80,0.87] / 0.088 [0.070,0.107] / 15.8 [12.9,19.0] / 71.2 [67.3,74.8]	0.78 [0.73,0.83] / 0.74 [0.68,0.80] / 0.70 [0.63,0.76]
BM25 + BioLinkBERT rerank + GBDT	0.63 [0.59,0.67] / 0.55 [0.51,0.59] / 0.59 [0.55,0.63]	0.792 [0.76,0.82] / 0.119 [0.098,0.142] / 20.7 [17.6,24.0] / 100.0	0.69 [0.63,0.75] / 0.64 [0.57,0.70] / 0.56 [0.49,0.63]

Table S28: Document grounding, raw decision/coverage counts, and five-seed σ_{EP} on the cross-rubric evaluation.

System	DocEP/DocER/DocF1	Correct/incorrect	Covered/withheld	σ_{EP}
GPT-4.1 + RAG + Sel.	0.80 [0.76,0.83] / 0.73 [0.69,0.77] / 0.76 [0.72,0.79]	393/51	444/156	0.0061
Claude-3.5 + RAG + Sel.	0.78 [0.74,0.81] / 0.71 [0.67,0.75] / 0.74 [0.70,0.77]	384/55	439/161	0.0065
Qwen-2.5-72B + RAG + Sel.	0.75 [0.71,0.79] / 0.68 [0.64,0.72] / 0.71 [0.67,0.75]	372/62	434/166	0.0073
Mixtral-8x22B + RAG + Sel.	0.722 [0.68,0.76] / 0.651 [0.61,0.69] / 0.684 [0.64,0.72]	364/67	431/169	0.0076
MedFound-34B + RAG + Sel.	0.73 [0.69,0.77] / 0.66 [0.62,0.70] / 0.69 [0.65,0.73]	366/66	432/168	0.0080
Llama-3.1 70B + RAG + Sel.	0.71 [0.67,0.75] / 0.64 [0.60,0.68] / 0.67 [0.63,0.71]	358/69	427/173	0.0077
BM25 + BioLinkBERT rerank + GBDT	0.57 [0.52,0.62] / 0.49 [0.44,0.54] / 0.53 [0.48,0.58]	475/125	600/0	–

Table S29: Five-seed execution variance on the 3,212-case frozen test. The top-three order is preserved in every run, with $\sigma_{\mathrm{EP}} \leq 0.0063$ and $\sigma_{\mathrm{OC}} \leq 0.51$ percentage points; paired system contrasts are reported in Table S30.

System	Mean EP	σ_{EP}	Mean OC (%)	σ_{OC} (pp)
GPT-4.1 + RAG + Sel.	0.928	0.0042	6.32	0.38
Claude-3.5 + RAG + Sel.	0.918	0.0047	6.81	0.44
Llama-3.1 70B + RAG + Sel.	0.858	0.0063	11.22	0.51

Table S30: Paired focal contrasts on the three-family frozen-test endpoint. Effect bundles report ΔEP [95% CI] / ΔEF1 [95% CI] / ΔOC [95% CI] / Δcoverage in percentage points. Family effects are eligibility/endpoint/safety ΔEP.

Contrast	Pooled paired effects	Family ΔEP	Seed-wise sign
GPT prompt → RAG	+0.117 [0.101,0.133] / +0.154 [0.136,0.172] / -5.7 [-7.6,-3.9] / 0.0	+0.11 / +0.13 / +0.14	5/5
Claude prompt → RAG	+0.120 [0.103,0.137] / +0.158 [0.139,0.177] / -6.0 [-8.0,-4.1] / 0.0	+0.11 / +0.14 / +0.15	5/5
Llama prompt → RAG	+0.123 [0.104,0.142] / +0.153 [0.132,0.174] / -7.0 [-9.2,-4.9] / 0.0	+0.12 / +0.13 / +0.12	5/5
GPT RAG → selective	+0.051 [0.036,0.066] / +0.059 [0.043,0.075] / -8.6 [-10.4,-6.9] / -20.3	+0.04 / +0.05 / +0.06	5/5
Claude RAG → selective	+0.058 [0.042,0.074] / +0.070 [0.053,0.087] / -8.9 [-10.8,-7.1] / -21.8	+0.05 / +0.05 / +0.07	5/5
Llama RAG → selective	+0.084 [0.064,0.104] / +0.097 [0.075,0.119] / -8.3 [-10.5,-6.2] / -25.0	+0.09 / +0.08 / +0.09	5/5
GPT selective − Claude selective	+0.014 [0.005,0.023] / +0.010 [0.002,0.018] / -0.6 [-1.8,0.5] / +1.5	+0.01 / +0.02 / +0.01	5/5
Claude selective − Llama selective	+0.061 [0.046,0.076] / +0.070 [0.053,0.087] / -4.5 [-6.1,-2.9] / +3.2	+0.04 / +0.05 / +0.08	5/5

K.1 Paired focal effects

Prompt-to-RAG effects pair all frozen-test cases because both rows have full coverage. RAG-to-selective effects restrict the full-coverage RAG row to the cases issued by its matched selective row; the final bundle component separately records the overall coverage change. Between-selective effects in Table S32 use each pair's joint-issued population. Thus every paired interval is computed on a common case set rather than by subtracting independently covered row averages.

K.2 Matched-case comparability audits

Two paired document audits close the population-matching loop. The regime audit samples 160 cases issued by all three variants of each backbone, for 480 case triplets and 1,440 reviewed outputs. The row-ordering audit samples 320 cases jointly issued by GPT, Claude, Qwen, Mixtral, MedFound, and Llama selective rows; the same cases are scored by the full-coverage BioLinkBERT rerank, BioLinkBERT, PubMedBERT, and Classical-hybrid controls, yielding 3,200 reviewed outputs. Cases are stratified by the three grounded families and therapeutic area; system identity and regime are hidden from the three clinicians. DocEP/DocF_1 differences use 10,000 paired stratified-bootstrap resamples, and win/tie/loss counts are computed casewise under the binary document-grounding rule.

Table S31: Matched-case document effects for the six within-backbone regime transitions. Values are ΔDocEP / ΔDocF_1 [paired 95% interval]; document-grounding outcomes are wins/ties/losses for the stronger regime.

Transition	Paired n	ΔDocEP / ΔDocF_1	W/T/L
GPT prompt → RAG	160	+0.110 [0.075,0.145] / +0.160 [0.121,0.199]	29/120/11
Claude prompt → RAG	160	+0.130 [0.093,0.167] / +0.160 [0.119,0.201]	32/117/11
Llama prompt → RAG	160	+0.130 [0.090,0.170] / +0.150 [0.106,0.194]	31/119/10
GPT RAG → selective	160	+0.100 [0.068,0.132] / +0.090 [0.055,0.125]	24/128/8
Claude RAG → selective	160	+0.070 [0.038,0.102] / +0.090 [0.055,0.125]	20/131/9
Llama RAG → selective	160	+0.090 [0.052,0.128] / +0.110 [0.069,0.151]	22/130/8

The matched populations preserve all 18 regime-transition metric claims, six direct selective-ordering pairs, and all three transparent-control pairs used by the final ledger. Qwen–Mixtral remains surrogate-only because both paired intervals cross zero; Mixtral–MedFound receives a reversal flag because the stronger document and cross-rubric audits change sign.

Table S32: Pairwise comparability for selective and transparent-control orderings. Automatic effects use every common primary-panel case; document effects use the shared 320-case blinded subset.

Comparison	Joint-issued n	ΔEP / ΔEF1	ΔDocEP / ΔDocF$_1$	Doc W/T/L
GPT − Claude	1,887	+0.013 [0.004,0.022] / +0.011 [0.003,0.019]	+0.022 [0.004,0.040] / +0.020 [0.003,0.038]	12/303/5
Claude − Qwen	1,834	+0.036 [0.022,0.051] / +0.035 [0.020,0.050]	+0.025 [0.007,0.044] / +0.020 [0.006,0.035]	15/298/7
Qwen − Mixtral	1,856	+0.009 [-0.004,0.022] / +0.010 [-0.005,0.025]	+0.006 [-0.016,0.028] / +0.008 [-0.015,0.031]	14/294/12
Mixtral − MedFound	1,807	+0.011 [-0.003,0.025] / +0.009 [-0.007,0.025]	-0.009 [-0.030,0.012] / -0.010 [-0.032,0.012]	10/297/13
MedFound − Llama	1,773	+0.006 [0.001,0.012] / +0.016 [0.005,0.027]	+0.022 [0.004,0.041] / +0.019 [0.002,0.036]	11/305/4
MedFound − BioLink rerank	1,932	+0.116 [0.101,0.131] / +0.153 [0.136,0.170]	+0.175 [0.136,0.214] / +0.199 [0.157,0.241]	68/240/12
Qwen − BioLink rerank	1,954	+0.137 [0.121,0.153] / +0.171 [0.153,0.189]	+0.216 [0.174,0.258] / +0.238 [0.193,0.283]	75/239/6
Llama − BioLink rerank	1,924	+0.113 [0.096,0.130] / +0.137 [0.118,0.156]	+0.150 [0.111,0.189] / +0.174 [0.132,0.216]	60/248/12
BioLink rerank − BioLink encoder	2,565	+0.041 [0.023,0.059] / +0.040 [0.021,0.059]	+0.059 [0.025,0.093] / +0.057 [0.021,0.093]	25/289/6
BioLink encoder − PubMedBERT	2,565	+0.039 [0.021,0.057] / +0.041 [0.022,0.060]	+0.053 [0.020,0.086] / +0.047 [0.013,0.081]	22/293/5
PubMedBERT − Classical hybrid	2,565	+0.087 [0.068,0.106] / +0.067 [0.047,0.087]	+0.103 [0.064,0.142] / +0.087 [0.046,0.128]	38/277/5

L HUMAN-STUDY, ETHICS, AND GOVERNANCE DOCUMENTATION

ANNOTATOR PANEL AND CLINICIAN AUDITS

Four human-judgment groups contribute to the benchmark: (i) the 14-person benchmark annotator panel (median 8 years trial-operations experience), which produced the 12,847 gold labels and 4,907 double-annotated instances; (ii) the 5-person main-results clinician panel, which scored the same frozen-test partition used for Table S1 and is reported there only through a panel-level consensus anchor rather than as five separate leaderboard rows; (iii) the 3-person independent audit panel, which produced the 384-instance external audit; and (iv) the 3-person citation panel, which completed the 800-output and matched-case document audits. Groups (i) and (ii) are disjoint; group (iii) is disjoint from both. Group (iv) is disjoint from the other panels and produced the document audits reported in Sections F and K.2.

WORKLOAD ACCOUNTING

The annotation-contract row and the clinician-audit rows cover distinct work packages. The 14×920 benchmark-annotation hours form the complete vendor-contract ledger across the benchmark-construction window: primary labeling for 12,847 instances, evidence-span marking, the 4,907 double-annotation pass, adjudication of the 8.4% disagreement subset, and post-adjudication quality control under the locked manual. The 5×120 clinician-panel hours cover manual familiarization, independent scoring for the frozen-test panel anchor, panel-consensus logging for the tabulated clinician row in Table S1, and evidence-boundary review on difficult cases. The separate 3×64 and 3×104 rows correspond to the prospectively frozen 384-instance independent audit and the 800-output document-level citation audit, respectively. The row-stratified extensions add 3×120 hours for 960 outputs. The additional 3×500 hours reuse the citation panel and locked rubric for 4,640 matched-case output judgments.

PARTICIPANT INSTRUCTIONS (SUMMARY)

Clinicians were given a locked 22-page scoring manual that specifies (i) task-family definitions, (ii) evidence-boundary rules, (iii) the 0–3 decision-sufficiency Likert scale used for the citation audit, (iv) adjudication escalation paths, and (v) data-handling requirements (no copying of any source text

Table S33: Human-study protocol and workload accounting. All clinicians were board-certified and compensated at $150/hour; WCG IRB classified the public-record study as non-human-subjects research (determination 2025-11-08).

Panel	Clinicians (n)	Hours per clinician	Compensation (USD)	Adjudication
Benchmark annotators (vendor contract)	14	920	$1,932,000 (annotation contract)	Gold-label creation, evidence-span marking, double annotation on 4,907 instances; senior adjudication for 8.4% disagreements
Main clinician panel (5 experts)	5	120	$90,000	Frozen-test panel anchor; independent per-clinician scoring, then recorded panel-consensus row for Table S1
Independent 384-instance audit	3	64	$28,800	Independent scoring; majority vote; $\kappa = 0.82$ with model
Citation-validity audit (800 outputs)	3	104	$46,800	Double-scoring on 60% of outputs; $\kappa = 0.79$ on binary indicator
Row-stratified audit extensions (960 outputs)	3 (same panel)	120	$54,000	Five added 160-output system slices plus the separate 160-output Classical-hybrid addendum
Matched-case document audits (4,640 outputs)	3 (same panel)	500	$225,000	Six regime contrasts and eleven row-ordering comparisons on common cases

outside the secure review portal). No clinician had access to system identities or to the results of any other panel during scoring. Scoring was performed asynchronously on a web portal that enforced single-instance viewing and time-stamped every rating. For the 5-clinician frozen-test anchor, the portal first collected blinded individual decisions and then logged one panel-consensus output per case for the tabulated clinician row, rather than averaging five separate clinician score vectors. The instruction summary below is sufficient to interpret the reported human-study results without relying on any additional document.

LOCKED MANUAL EXCERPT (DECISION-CRITICAL RULES)

The decision-critical manual rules are: (i) assign the task label before marking the minimal supporting evidence; (ii) flag and escalate cases where two labels remain defensible; (iii) score citation sufficiency as 0 = unsupported, 1 = partially relevant, 2 = sufficient with minor omissions, and 3 = fully sufficient; and (iv) do not infer unstated clinical facts or use knowledge absent from the task record.

RISK DISCLOSURE

Participants were informed that (i) all source data are public records, (ii) no identifiable participant information is surfaced in task payloads, (iii) their ratings would be reported only as aggregate metrics in the paper, and (iv) they could withdraw at any time without penalty. Individual rating logs are not reported; only aggregate agreement and decision-sufficiency distributions appear in the submission.

Table S34: Source licenses and terms for every public record collection used by ADMITSCOPE.

Asset	Upstream license / terms	Use in ADMITSCOPE
ClinicalTrials.gov protocol records	ClinicalTrials.gov terms (U.S. government database; possible third-party copyright in contributed records) ClinicalTrials.gov (2023)	Derived eligibility, endpoint, and RBM task fields
FAERS narratives	U.S. public domain	Safety signal detection and endpoint cases
ISRCTN records (UK)	ISRCTN terms (post-2019 metadata CC0; individually marked open-access content under its stated licence) ISRCTN (n.d.)	External-validity slice (eligibility, endpoint)
WHO ICTRP trial records	WHO ICTRP terms (public download; attribution, processing-date, and non-commercial-use conditions) World Health Organization (n.d.)	External-validity slice (eligibility)
EudraVigilance public release	EMA EudraVigilance access policy (restricted public data elements) European Medicines Agency (2025)	External-validity slice (safety)
UK CRF monitoring exemplars	UK Open Government Licence v3.0	RBM and UK CRF external-validity

M COMPUTATIONAL RESOURCES

N DRIFT AND INDEPENDENT-AUDIT NUMERIC SUMMARY

This section collects every temporal-drift and 384-instance independent-clinician-audit number asserted in the main paper into a single numeric display item, so that the main-text prose numbers are traceable to an appendix table rather than only to a figure or running text.

Table S35: Data-governance controls defining admitted sources, sensitive-information handling, intended evaluation use, schema stability, and refresh policy.

Governance control	Value
Source restriction	Only public ClinicalTrials.gov, FAERS, ISRCTN, WHO ICTRP, EudraVigilance, and UK CRF records that satisfy the locked inclusion rules are admitted.
Sensitive-information policy	Non-public free text is excluded; no direct participant identifiers are surfaced in task payloads (Table S6).
Schema stability	Task schema, split membership, and scorer configuration are locked for this submission and summarized in the tables and ledgers cited throughout the appendix.
Intended use	Comparative evaluation of evidence-bearing systems on public-source trial-operations tasks.
Evaluated scope	Public protocol, safety, and monitoring records; transformed-input RBM stress; source-native external and cross-rubric evaluation (Table S5).
Licensed-use boundary	Claim licenses apply to comparative evaluation in the documented public-record settings and evaluated workflow paths.
Refresh policy	Any future refresh must preserve explicit split accounting, source provenance, and drift reporting against earlier versions.
Safety controls	Safety narratives use public FAERS records and already-public attributes; clinician re-reading is used for the strongest citation audits.

Table S36: Computational resources used for every experiment reported in the paper, including failed and preliminary runs. All GPU measurements were taken on a compute cluster of NVIDIA A100 80GB nodes (CUDA 12.1, driver 535.129.03). API-hosted backbones were called using the provider-side snapshot tags listed (`gpt-4.1-2025-04-14`, `claude-3-5-sonnet-20241022`). Retrieval used FAISS 1.7.4 HNSW32 over 428,517 passages; BM25 used the Pyserini-0.22.1 implementation. The software stack was Python 3.11.7, NumPy 1.26.3, PyTorch 2.2.2, and vLLM 0.5.4. The table is intended as a complete experiment ledger rather than as a headline cost claim.

Component	Hardware / API	Reported experiments (incl. failed runs)
GPT-4.1 inference (all runs)	API, snapshot `gpt-4.1-2025-04-14`	41,240 requests, 52.4 API-hours (main), +18 API-hours (variance + external slices)
Claude-3.5 inference (all runs)	API, snapshot `claude-3-5-sonnet-20241022`	41,240 requests, 48.1 API-hours (main), +17 API-hours (variance + external slices)
Llama-3.1 70B inference (all runs)	4× A100 80GB, vLLM 0.5.4	62.3 GPU-hours (main + variance + external slices)
MedFound-34B inference	2× A100 80GB	14.1 GPU-hours
FAISS index build + retrieval	1× A100 80GB	5.9 GPU-hours (build); ≈12 ms / query p50 serving
Classical hybrid (BM25+GBDT)	AMD EPYC 7763 CPU node (64 cores)	22.4 CPU-hours (train); 3.1 CPU-hours (inference)
Citation-validity audit tooling	CPU + web portal	14.0 CPU-hours
Failed / preliminary experiments	Mixed	76 GPU-hours and 63 API-hours accumulated during freeze-before-report testing
Total project compute	–	≈ 158 GPU-hours, ≈ 198 API-hours, ≈ 40 CPU-hours

Table S37: Temporal-drift panel on the 1,843-instance out-of-main-pool slice and independent 384-instance prospectively frozen clinician-audit subset, plus the 412-site-month RBM anchor used in Supplementary Figure S4. Temporal-drift rows compare the frozen-test reference to the temporal slice; ΔEP and ΔOC are expressed as percentage-point shifts. The mean ΔEP across evaluated systems on the temporal slice is 2.4 pp and the mean ΔOC is 2.1 pp, matching the values reported in Section 5.3. The 384-instance audit rows give the clinician-panel and GPT-4.1 + RAG + selective numeric anchors stated in the main-paper Independent Clinician Audit paragraph (panel accuracy 0.963, panel EP 0.951; GPT-4.1 accuracy 0.941, EP 0.914, coverage 81.0%, model–panel $\kappa = 0.82$). The RBM-anchor row reports Cohen's $\kappa = 0.73$ agreement between RBM gold tiers and independent public monitoring-exemplar escalations across 412 site-month panels.

Panel / slice	Metric A	Metric B	Metric C	ΔEP (pp)	ΔOC (pp)
Temporal slice ($n = 1{,}843$; 2025-04-01 to 2026-03-31; excluded from development)					
GPT-4.1 prompt-only (frozen $\rightarrow$ temporal)	EP 0.77 $\rightarrow$ 0.728	Acc 0.84 $\rightarrow$ 0.81	OC 19.8% $\rightarrow$ 22.7%	-4.2	$+2.9$
GPT-4.1 + RAG + Sel. (frozen $\rightarrow$ temporal)	EP 0.93 $\rightarrow$ 0.912	Acc 0.95 $\rightarrow$ 0.94	OC 6.3% $\rightarrow$ 8.4%	-1.8	$+2.1$
Claude-3.5 + RAG + Sel. (frozen $\rightarrow$ temporal)	EP 0.92 $\rightarrow$ 0.904	Acc 0.95 $\rightarrow$ 0.94	OC 6.8% $\rightarrow$ 8.4%	-1.6	$+1.6$
Llama-3.1 70B + RAG + Sel. (frozen $\rightarrow$ temporal)	EP 0.86 $\rightarrow$ 0.840	Acc 0.91 $\rightarrow$ 0.89	OC 11.2% $\rightarrow$ 13.0%	-2.0	$+1.8$
Cross-system mean	–	–	–	-2.4	$+2.1$
Independent 384-instance clinician audit (prospectively frozen; independent panel disjoint from benchmark annotators)					
Clinician panel ($n = 3$)	Acc 0.963	EP 0.951	Cov. 100%	–	–
GPT-4.1 + RAG + Sel.	Acc 0.941	EP 0.914	Cov. 81.0% ($\kappa = 0.82$)	–	–
RBM construction anchor (412 site-month panels, bounded-proxy validation; see Fig. S4)					
RBM gold tier vs. independent exemplar escalation	$\kappa = 0.73$	–	–	–	–

O OPEN-WEIGHT AND TASK-STRUCTURED BASELINES (SUPPLEMENT-REPORTED)

This appendix extends the primary roster with a fully open-weight Qwen-2.5-72B-Instruct configuration and a biomedical task-structured PubMedBERT + GBDT pipeline. Qwen is served with vLLM 0.5.4 on $4\times$ NVIDIA A100 80GB (CUDA 12.1, driver 535.129.03) and uses the same FAISS HNSW32 retrieval index, temperature scaling, and split-conformal abstention policy as the proprietary selective systems. PubMedBERT is frozen, paired with four family-specific GBDT heads trained on the corresponding family partitions of the 7,708-case training split, and retrieves from the same 428,517-passage corpus. The Qwen configuration is evaluated across five fixed prompt-ordering and vLLM seeds at `temperature = 0, top_p = 1.0`.

Table S38: Open-weight and task-structured baselines on the 3,212-case frozen test. Qwen selective reaches 0.890 EP across five seeds, and PubMedBERT + GBDT improves the classical BM25 + GBDT floor by 0.08 EP.

Configuration	Mean EP	σ_{EP}	Mean EF1	Mean Acc.	Mean OC (%)	Mean Cov. (%)
Qwen-2.5-72B-Instruct prompt	0.712	0.0074	0.61	0.80	22.8	100.0
Qwen-2.5-72B-Instruct + RAG	0.843	0.0071	0.78	0.87	17.1	100.0
Qwen-2.5-72B-Instruct + RAG + Sel.	0.890	0.0059	0.86	0.93	9.2	75.9
PubMedBERT + GBDT (task-structured)	0.650	–	0.58	0.78	23.4	100.0

EXPANDED OPEN AND TRANSPARENT PANEL

We add a second open-weight selective model (Mixtral-8x22B-Instruct) and a second transparent structured baseline (BioLinkBERT + family-specific GBDT head), both under the same retrieval index, frozen prompts, calibration pool, and scorer. Frozen-test results use the same 3,212-case partition; stochastic selective rows use five fixed seeds, structured rows are deterministic, and intervals follow the 10,000-resample stratified-bootstrap convention. Metric bundles report EP/EF1/accuracy/OC/coverage.

Table S39: Expanded open and transparent panel on 3,212 frozen-test cases. Selective rows report five-seed means, intervals, and issued/withheld counts; deterministic structured rows use full coverage.

Configuration	Metrics	ER/ECE	95% CI (EP/EF1/Cov.)	Issued/withheld	3-family EP/EF1/OC	σ_{EP}
Qwen-2.5-72B-Instruct + RAG + Sel.	0.890 / 0.86 / 0.93 / 9.2 / 75.9	0.842 / 0.074	0.872,0.907 / 0.841,0.877 / 73.3,78.4	2438/774	0.881 / 0.851 / 8.8	0.0059
Mixtral-8x22B-Instruct + RAG + Sel.	0.884 / 0.851 / 0.91 / 9.6 / 76.8	0.835 / 0.081	0.864,0.902 / 0.831,0.870 / 74.1,79.3	2467/745	0.873 / 0.842 / 9.2	0.0064
PubMedBERT + GBDT (task-structured)	0.650 / 0.58 / 0.78 / 23.4 / 100.0	0.531 / 0.149	0.624,0.676 / 0.555,0.604 / full	full	0.662 / 0.596 / 22.1	–
BioLinkBERT + GBDT (task-structured)	0.688 / 0.621 / 0.79 / 21.9 / 100.0	0.566 / 0.138	0.663,0.713 / 0.598,0.644 / full	full	0.701 / 0.637 / 20.4	–

LOCKED-INTERFACE ADDENDUM

The locked-interface addendum evaluates a selective MedFound-34B row and a BM25 + BioLinkBERT rerank + GBDT row under the same FAISS index, prompts, calibration pool, abstention rule, and scorer as the remaining panel. Metric bundles report EP/EF1/accuracy/OC/coverage.

Table S40: Locked-interface open and transparent addendum. MedFound tests a second biomedical selective model, while BM25 + BioLinkBERT rerank + GBDT strengthens the transparent retrieval control.

Configuration	Metrics	ER/ECE	95% CI (EP/EF1/Cov.)	Issued/withheld	3-family EP/EF1/OC	σ_{EP}
MedFound-34B + RAG + Sel.	0.871 / 0.842 / 0.90 / 10.7 / 75.1	0.824 / 0.086	0.852,0.888 / 0.821,0.862 / 72.5,77.6	2411/801	0.861 / 0.833 / 10.1	0.0071
BM25 + BioLinkBERT rerank + GBDT	0.731 / 0.661 / 0.81 / 18.6 / 100.0	0.603 / 0.123	0.708,0.752 / 0.637,0.684 / full	full	0.742 / 0.677 / 17.4	–

ADDENDUM PROTOCOL NOTES

MedFound-34B + RAG + selective uses the same $2\times$ A100 80GB serving stack already reported in Table S36, with retrieval traces frozen to the shared FAISS HNSW32 index and five fixed prompt-ordering / decoder seeds. The BM25 + BioLinkBERT rerank + GBDT baseline is fully transparent: BM25 produces the candidate set, a BioLinkBERT passage scorer reranks the top 32 retrieved passages on $1\times$ A100 80GB, and a family-specific GBDT head emits the decision and confidence using the same evidence passages and abstention interface as the other transparent baselines.

SEED-LEVEL QWEN-2.5-72B-INSTRUCT + RAG + SEL. TRAJECTORY

The five seed values for the strongest open-weight configuration are $\mathrm{EP} = \{0.890, 0.882, 0.897, 0.887, 0.894\}$ (mean 0.890, $\sigma_{\mathrm{EP}} = 0.0059$) with $\mathrm{OC} = \{9.2, 9.6, 8.9, 9.3, 9.0\}\%$ (mean 9.2%, $\sigma_{\mathrm{OC}} = 0.27$ pp). These seeds were fixed before test-set scoring.

OPEN-WEIGHT COMPUTE ACCOUNTING

Qwen-2.5-72B-Instruct inference ran on $4\times$ NVIDIA A100 80GB nodes (CUDA 12.1, driver 535.129.03) with vLLM 0.5.4; total wall time across the main run and five-seed variance study is 34.7 GPU-hours. PubMedBERT + GBDT encoder extraction required 6.2 GPU-hours on $1\times$ A100 80GB; GBDT training used the AMD EPYC 7763 CPU node listed in Table S36 for 4.1 CPU-hours. The two compute tables cover disjoint ledgers: the ≈ 158 GPU-hours / ≈ 198 API-hours total in Table S36 covers the 11 main-roster configurations, while Table S41 adds 40.9 GPU-hours and 4.1 CPU-hours for the supplement-reported open-weight and task-structured baselines.

Table S41: Incremental compute ledger for the supplement-reported baselines introduced in this section (added to the totals reported in Table S36). All hardware was drawn from the same reported compute cluster, and both baselines execute locally under the listed stacks. The Qwen row covers the main run plus the five-seed variance package reported in Table S38. The PubMedBERT rows separate encoder extraction from the family-specific GBDT heads so readers can see exactly where GPU versus CPU time is spent. These costs are incremental to the main-paper ledger rather than a second incompatible accounting. The table therefore documents the reproducibility burden of the supplement-reported baselines without changing any headline totals.

Component	Hardware	Reported experiments (incl. five-seed variance)
Qwen-2.5-72B-Instruct inference	$4\times$ A100 80GB, vLLM 0.5.4, CUDA 12.1	Main run + five seeds; 34.7 GPU-hours; ≈41,200 requests
PubMedBERT encoder extraction	$1\times$ A100 80GB	Frozen encoder over 428,517 passages; 6.2 GPU-hours
Family-specific GBDT heads (4 families)	AMD EPYC 7763 CPU node (64 cores)	Train on each family's partition of 7,708 total instances; 4.1 CPU-hours total

P CONSTRUCT-VALIDITY AUDIT UNDER PIPELINE-CUE WEAKENING

AUDIT DESIGN

A 312-instance balanced stratified audit (78 per family) was constructed to test whether the benchmark's qualitative ordering depends on surface cues introduced by the construction pipeline. For every sampled instance, the *same underlying source case* was re-rendered under three modifications, locked before any re-scoring: (i) author-authored *templated* text (canonical rule templates for eligibility, (metric, threshold, window, resolution) tuples for endpoint, locked safety-rubric framings, and RBM trigger-gate phrasings) was removed and replaced with a plain restatement of the raw source field; (ii) remaining source fields (e.g., criterion lists, FAERS narrative segments, site-metric entries) were randomly reordered under a seed locked per instance before re-scoring and recorded in the construct-validity audit ledger; (iii) the natural-language framing was independently paraphrased by an external medical-editing service under a locked rubric that preserves every clinical fact and cited span verbatim while paraphrasing connective text. Gold labels and gold evidence offsets were locked on the re-rendered payloads by a two-clinician adjudication pass (inter-annotator $\kappa = 0.86$; 9 adjudicated disagreements). We reran the prompt-only, RAG, and RAG + selective variants for GPT-4.1, Claude-3.5, and Llama-3.1 70B; the Qwen-2.5-72B selective row and the BioLinkBERT-rerank and Classical-hybrid controls were evaluated under the same weakened cues. Every configuration retained its main-run prompt, retrieval index, calibration parameters, and scorer settings; no retuning was permitted.

Table S42: Construct-validity audit on the 312-instance pipeline-cue-weakened subset. Absolute EP dips modestly for all three top selective systems, but the prompt $<$ RAG $<$ RAG + selective ordering is preserved within each backbone, the top-3 between-system ranking is preserved in all five seeds, and the three-family (RBM-excluded) pattern is preserved. Degradation is smallest on eligibility and endpoint adjudication (the rule-heavy families) and largest on RBM (the bounded-proxy family), consistent with the expectation that derived/templated content matters more for RBM than for the three directly-grounded families. All intervals are 95% stratified bootstrap confidence intervals over instances. The locked same-case intervention isolates dependence on construction-pipeline surface cues while preserving the source facts and evidence spans.

System / slice	EP	EF1	Elig. ΔEP	Safety ΔEP	RBM ΔEP
Frozen test → construct-validity audit (312 re-rendered cases, pipeline cues weakened)					
GPT-4.1 prompt-only	0.71 [0.67, 0.74]	0.63 [0.60, 0.67]	−0.05	−0.07	−0.11
GPT-4.1 + RAG	0.83 [0.80, 0.86]	0.78 [0.75, 0.81]	−0.04	−0.06	−0.10
GPT-4.1 + RAG + Sel. (cov. 78.1%)	0.88 [0.85, 0.90]	0.85 [0.82, 0.88]	−0.04	−0.05	−0.12
Claude-3.5 prompt-only	0.69 [0.65,0.73]	0.61 [0.57,0.65]	-0.05	-0.08	-0.12
Claude-3.5 + RAG	0.81 [0.77,0.84]	0.75 [0.71,0.79]	-0.04	-0.07	-0.11
Claude-3.5 + RAG + Sel. (cov. 76.7%)	0.87 [0.84, 0.89]	0.84 [0.81, 0.87]	−0.04	−0.05	−0.12
Llama-3.1 70B prompt	0.60 [0.56,0.64]	0.52 [0.48,0.56]	-0.05	-0.08	-0.12
Llama-3.1 70B + RAG	0.72 [0.68,0.76]	0.66 [0.62,0.70]	-0.04	-0.07	-0.11
Llama-3.1 70B + RAG + Sel. (cov. 73.9%)	0.81 [0.77, 0.84]	0.77 [0.74, 0.80]	−0.04	−0.05	−0.11
Classical hybrid (BM25+GBDT)	0.51 [0.47, 0.54]	0.47 [0.44, 0.50]	−0.06	−0.07	−0.06

INTERPRETATION

The cue intervention preserves the within-backbone prompt $<$ RAG $<$ RAG + selective ordering on 3/3 backbones and the GPT $>$ Claude $>$ Llama $>$ classical ordering on the aggregate and every

family. For GPT-4.1 + RAG + selective, ΔEP is -0.04 on eligibility, -0.05 on endpoint, -0.05 on safety, and -0.12 on RBM; Claude-3.5 and Llama-3.1 reproduce the same family profile. The regime gap widens on the three grounded families, showing that retrieval and calibrated selection retain more of their advantage after surface cues are weakened. This same-case intervention therefore grants direct cue-robust support to the grounded-family regime claims and identifies transformed RBM inputs as the strongest cue-sensitivity stress.

Independent cue-weakening upgrade

To reduce author mediation, we add a third-party source-native rewrite and citation-decoy audit on 360 grounded cases balanced as 120/120/120 across eligibility / endpoint / safety. An external medical-editing vendor rewrites connective text under a locked fact-preservation rubric, a deterministic audit script injects decoy citations before blinded adjudication, and prompts, retrieval, calibration, and thresholds remain frozen from the main calibration pool. We report 95% stratified-bootstrap intervals together with raw label-support and evidence-support failure counts, because the strongest threat here is not mere accuracy loss but successful exploitation of benchmark-surface cues.

Table S43: Independent 360-case cue-weakening and citation-decoy audit on the directly grounded families. The selective order remains intact, and evidence-support failures outnumber label failures for every system.

System	EP/EF1	OC/Cov.	Label / evidence failures
GPT-4.1 + RAG + Sel.	0.86 [0.82,0.89] / 0.83 [0.79,0.86]	9.8 [7.6,12.4] / 75.1 [70.8,79.0]	18/32
Claude-3.5 + RAG + Sel.	0.85 [0.81,0.88] / 0.82 [0.78,0.85]	10.4 [8.0,13.1] / 73.8 [69.3,77.9]	20/35
Qwen-2.5-72B + RAG + Sel.	0.82 [0.78,0.86] / 0.79 [0.75,0.82]	11.1 [8.7,13.9] / 72.7 [68.2,76.8]	24/39
Llama-3.1 70B + RAG + Sel.	0.80 [0.76,0.84] / 0.76 [0.72,0.80]	13.8 [11.1,16.9] / 71.0 [66.4,75.3]	28/46

Q Post-cutoff contamination-robust slice

Slice construction

A 487-instance post-cutoff slice was constructed from public records registered between 2025-10-15 and 2026-03-31, after every provider-declared or model-card-stated cutoff in Table S44. The slice contains 158 ClinicalTrials.gov protocols, 172 FAERS 2025Q4/2026Q1 narratives, 81 UK CRF 2025H2 monitoring exemplars, and 76 EudraVigilance 2025Q4–2026Q1 public notifications. No instance shares an NCT ID, FAERS case number, UK CRF exemplar identifier, or ICTRP UUID with the 12,847-instance main pool, the 1,843-instance temporal slice, or the 612-instance external slice. The slice was frozen on 2026-04-05; prompts, retrieval configuration, calibration parameters, and scoring remained fixed.

Reconciliation with other robustness views

The post-cutoff drop has the same direction as the 5–7-point external-slice shift and is substantially larger than the 0.3–0.4-point clean/full sensitivity. The post-cutoff window is the stronger temporal test because every record follows the documented model cutoffs; the stable top-three order provides direct evidence that the prompt $<$ RAG $<$ selective separation persists beyond observable public-web overlap.

R Admissibility evidence matrix

The complete matrix instantiates five stress models at row level: citation-overlap gaming, prompt-surface dependence, post-cutoff contamination, independent source shift, and selective-coverage mismatch. Each row records the protected comparison, activated obligation, strongest comparator,

Table S44: Model snapshots and source-documented knowledge cutoffs relative to the post-cutoff evaluation window.

Backbone snapshot	Documented cutoff	Evaluation window	Primary source
`gpt-4.1-2025-04-14`	2024-06-01	2025-10-15–2026-03-31	OpenAI model documentation OpenAI (2025)
`claude-3-5-sonnet-20241022`	April 2024	2025-10-15–2026-03-31	Anthropic model-card addendum Anthropic (2024b)
`llama-3.1-70b-instruct`	December 2023	2025-10-15–2026-03-31	Meta Llama 3.1 model card Meta (2024)

Table S45: Post-cutoff re-evaluation on 487 records registered after every documented model cutoff. The top-three order persists after an 8–9-point EP shift, and each selective row remains above the classical hybrid.

System (slice)	EP	EF1	Acc.	OC (%)	Cov. (%)
GPT-4.1 prompt-only (post-cutoff)	0.66 [0.62, 0.70]	0.59 [0.55, 0.63]	0.78 [0.74, 0.81]	24.8 [21.1, 28.7]	100.0
GPT-4.1 + RAG (post-cutoff)	0.79 [0.75, 0.82]	0.74 [0.70, 0.78]	0.85 [0.82, 0.88]	18.4 [15.3, 21.8]	100.0
GPT-4.1 + RAG + Sel. (post-cutoff)	0.85 [0.82, 0.88]	0.82 [0.78, 0.85]	0.90 [0.87, 0.92]	10.6 [8.4, 13.1]	73.2 [69.3, 77.0]
Claude-3.5 + RAG + Sel. (post-cutoff)	0.84 [0.80, 0.87]	0.81 [0.77, 0.84]	0.89 [0.86, 0.91]	11.0 [8.7, 13.6]	72.1 [68.2, 75.9]
Llama-3.1 70B + RAG + Sel. (post-cutoff)	0.77 [0.73, 0.80]	0.73 [0.69, 0.76]	0.85 [0.82, 0.88]	14.9 [12.2, 17.9]	69.4 [65.4, 73.3]
Classical hybrid (post-cutoff)	0.49 [0.45, 0.53]	0.45 [0.41, 0.49]	0.69 [0.65, 0.73]	32.4 [28.7, 36.2]	100.0
GPT-4.1+RAG+Sel. Δ (frozen → post-cutoff)	−0.08	−0.08	−0.05	+4.3 pp	−6.2 pp
Claude-3.5+RAG+Sel. Δ	−0.08	−0.08	−0.06	+4.2 pp	−5.8 pp
Llama-3.1 70B+RAG+Sel. Δ	−0.09	−0.09	−0.06	+3.7 pp	−5.4 pp

held-out test, support license, utility cost, observed boundary, and source trace. This is the fully expanded compiler output behind the main-paper admissibility frontier.

Table S46: Complete system-row evidence matrix for all evaluated configurations, including comparator, held-out setting, tier, utility cost, observed failure mode, and evidence trace.

Protected row	Property	Stress model	Trust assumption	Strongest comparator	Held-out / adaptive setting	Evidence tier	Utility / overhead	Observed boundary	Trace
Clinician panel	Calibration	Panel variance	Consensus anchor is stable enough for calibration reference	Top selective rows	Frozen test only	Human anchor	100% cov.; clinician-hours only	Calibration anchor only	Table S1; Table S33
GPT-4.1 prompt-only	Grounding	Coverage-insensitive analysis	Automatic scorer tracks evidence failures coarsely	GPT-4.1 + RAG / +Sel.	Paired Doc + post-cutoff	Direct regime contrast	Full coverage; no abstention overhead	Lowest document support among matched GPT rows	Table S1; Tables S31, S45
Claude-3.5 prompt-only	Grounding	Coverage-insensitive analysis	Same prompt-only interface is a valid control	Claude-3.5 + RAG / +Sel.	Paired Doc	Direct regime contrast	Full coverage	Paired-document scope	Table S1; Table S31

Protected row	Property	Stress model	Trust assumption	Strongest comparator	Held-out / adaptive setting	Evidence tier	Utility / overhead	Observed boundary	Trace
Llama-3.1 70B prompt	Grounding	Coverage-insensitive analysis	Prompt-only open row remains a valid lower-band control	Llama-3.1 70B + RAG / +Sel.	Paired Doc	Direct regime contrast	Full coverage	Paired-document scope	Table S1; Table S31
MedFound-34B prompt	Grounding	Open-roster stress	Prompt-only medical open row is a meaningful open control	MedFound + RAG + Sel.	Frozen test	Span-only open control	Full coverage	Frozen-test surrogate scope	Table S1; Table S40
GPT-4.1 + RAG	Grounding	Retrieval-null stress	Retrieval gain is not pure cue matching	GPT-4.1 prompt / +Sel.	Paired Doc + post-cutoff	Direct regime contrast	Full coverage; retrieval trace logging	Off-source audit limited to post-cutoff records	Table S1; Tables S31, S45
Claude-3.5 + RAG	Grounding	Trust-transfer stress	Mid-tier RAG rows lose more under human review	Claude-3.5 + RAG + Sel.	Widened + paired Doc	Direct regime contrast	Full coverage; focused re-read	No second independent workflow path	Table S1; Tables S21, S31
Llama-3.1 70B + RAG	Grounding	Retrieval-null stress	Open-backbone RAG provides a matched regime contrast	Llama prompt / +Sel.	Paired Doc	Direct regime contrast	Full coverage	Paired-document scope	Table S1; Table S31
Classical hybrid (BM25 + GBDT)	Lower bound	Lower-bound stress	Transparent floor should remain visibly below selective rows	Open / proprietary selective rows	External, post-cutoff, and doc addendum	Doc audit + off-source lower bound	Full coverage; transparent retrieval	Evidence support remains weak under human review	Table S1; Table S22; Table S45
GPT-4.1 + RAG + Sel.	Ranking	Citation gaming + source shift	Top row remains trustworthy under stronger review	Claude-3.5 + RAG + Sel.	Joint-issued Doc, ext., dual holdout, cue, post-cutoff, variance	Paired Doc + indep.	79.4% cov.; focused clinician re-reading	Safety drives the largest off-source drop	Table S1; Tables S32, S25, S26
Claude-3.5 + RAG + Sel.	Ranking	Citation gaming + source shift	Ordering below GPT survives matched review	GPT-4.1 + RAG + Sel.	Joint-issued Doc, ext., dual holdout, cue, post-cutoff, variance	Paired Doc + indep.	77.9% cov.	Safety drives the largest off-source drop	Table S1; Tables S32, S25, S26
Llama-3.1 70B + RAG + Sel.	Ranking	Citation gaming + source shift	Open major-backbone row preserves selective ordering	MedFound + RAG + Sel.	Joint-issued Doc, ext., dual holdout, cue, post-cutoff, variance	Paired Doc + indep.	74.8% cov.	Larger absolute source-shift drop than top two rows	Table S1; Tables S32, S25, S26
Qwen-2.5-72B + RAG + Sel.	Ranking	Proprietary-only stress	Open selective row should survive partial transfer	Llama + Sel. / BioLink rerank	Joint-issued Doc, dual holdout, cue	Direct vs rerank; indep. transfer	75.9% cov.; five seeds	Qwen–Mixtral interval crosses zero	Tables S32, S25
Mixtral-8x22B + RAG + Sel.	Ranking	Small-margin stress	Second open selective row lands in the same broad band	Qwen / MedFound + RAG + Sel.	Joint-issued Doc + cross-rubric	Surrogate; reversal flag	76.8% cov.; five seeds	Small margin reverses against MedFound	Tables S32, S27
MedFound-34B + RAG + Sel.	Ranking	Proprietary-only stress	Medical-domain open row remains above transparent floors	Llama + Sel. / BioLink rerank	Joint-issued Doc + cross-rubric	Direct matched support	75.1% cov.; five seeds	Safety interval remains widest	Tables S32, S27
PubMedBERT + GBDT	Lower bound	Retrieval-floor stress	Task-structured non-LLM stack beats the BM25 floor	BM25 + GBDT	Joint-issued automatic + Doc	Direct matched support	Full coverage; deterministic inference	Frozen-test document scope	Tables S32, S39

Protected row	Property	Stress model	Trust assumption	Strongest comparator	Held-out / adaptive setting	Evidence tier	Utility / overhead	Observed boundary	Trace
BioLinkBERT + GBDT	Lower bound	Retrieval-floor stress	Stronger biomedical encoder improves the floor	PubMedBERT + GBDT	Joint-issued + automatic + Doc	Direct matched support	Full coverage; deterministic inference	Frozen-test document scope	Tables S32, S39
BM25 + BioLinkBERT rerank + GBDT	Lower bound	Retrieval-centric stress	Transparent reranking narrows the open-stack gap	BioLinkBERT + GBDT / MedFound + RAG + Sel.	Joint-issued + Doc + cross-rubric	Direct matched support	Full coverage; transparent top-32 rerank	Largest span-to-document gap	Tables S32, S21, S27

S COMPLETE CLAIM-SURVIVAL LEDGER

The ledger below enumerates the 48 predeclared metric-specific comparisons. The compiler fixes applicability from claim type, shared audit rubric, and comparison population before scoring. "Direct" requires a positive paired automatic margin and directional preservation under every active claim-matched audit; "transferred" records support from a neighboring predeclared held-out setting; "surrogate-only" binds the conclusion to the public scorer; and "reversed" flags a sign change under stronger review. Inapplicable checks provide neither support nor failure.

Table S47: Complete claim-survival ledger. Multi-ID rows enumerate one claim per named metric.

Claim IDs	Comparison and public margin	Audit applicability	Document evidence	Independent / temporal evidence	Final tier	Decision / source
C01 EP; C02 EF1; C03 OC	GPT prompt → RAG: $+.117/+.154/-5.7$	Matched backbone; full coverage	Paired ΔDocEP/DocF1 $+.110/+.160$	Five-seed paired intervals preserve all signs	Direct	Retain; Tables S30, S31
C04 EP; C05 EF1; C06 OC	Claude prompt → RAG: $+.120/+.158/-6.0$	Matched backbone; full coverage	Paired ΔDocEP/DocF1 $+.130/+.160$	Five-seed paired intervals preserve all signs	Direct	Retain; Tables S30, S31
C07 EP; C08 EF1; C09 OC	Llama prompt → RAG: $+.123/+.153/-7.0$	Matched backbone; full coverage	Paired ΔDocEP/DocF1 $+.130/+.150$	Five-seed paired intervals preserve all signs	Direct	Retain; Tables S30, S31
C10 EP; C11 EF1; C12 OC	GPT RAG → selective: $+.051/+.059/-8.6$	Matched retrieval; common issued cases	Paired ΔDocEP/DocF1 $+.100/+.090$	Five-seed signs $5/5$	Direct	Retain; Tables S30, S31
C13 EP; C14 EF1; C15 OC	Claude RAG → selective: $+.058/+.070/-8.9$	Matched retrieval; common issued cases	Paired ΔDocEP/DocF1 $+.070/+.090$	Five-seed signs $5/5$	Direct	Retain; Tables S30, S31
C16 EP; C17 EF1; C18 OC	Llama RAG → selective: $+.084/+.097/-8.3$	Matched retrieval; common issued cases	Paired ΔDocEP/DocF1 $+.090/+.110$	Five-seed signs $5/5$	Direct	Retain; Tables S30, S31
C19 EP; C20 EF1	GPT selective > Claude: paired $+.013/+.011$	Joint-issued primary cases	Paired DocEP/DocF1 $+.022/+.020$	Order retained on three paths	Direct	Retain; Tables S32, S27
C21 EP; C22 EF1	Claude selective > Qwen: paired $+.036/+.035$	Joint-issued primary cases	Paired DocEP/DocF1 $+.025/+.020$	Cross-rubric $+.03/+.03$	Direct	Retain; Tables S32, S27
C23 EP; C24 EF1	Qwen selective > Mixtral: paired $+.009/+.010$	Joint-issued primary cases	Paired intervals cross zero	Same ordering is not resolved independently	Surrogate-only	Downgrade; Table S32
C25 EP; C26 EF1	Mixtral selective > MedFound: paired $+.011/+.009$	Joint-issued primary and cross-rubric cases	Paired DocEP/DocF1 $-.009/-.010$	-0.005 [-0.018,0.008] / -0.008 [-0.022,0.006]	Surrogate-only; reversal	MedFound > Mixtral; Tables S32, S28
C27 EP; C28 EF1	MedFound selective > Llama: paired $+.006/+.016$	Joint-issued primary cases	Paired DocEP/DocF1 $+.022/+.019$	Cross-rubric $+.01/+.02$	Direct	Retain; Tables S32, S27
C29 EP; C30 EF1	MedFound selective > BioLink rerank: paired $+.116/+.153$	Joint-issued primary cases	Paired DocEP/DocF1 $+.175/+.199$	Cross-rubric direction retained	Direct	Retain; Tables S32, S27
C31 EP; C32 EF1	Qwen selective > BioLink rerank: paired $+.137/+.171$	Joint-issued primary cases	Paired DocEP/DocF1 $+.216/+.238$	Cross-rubric direction retained	Direct	Retain; Tables S32, S27
C33 EP; C34 EF1	Llama selective > BioLink rerank: paired $+.113/+.137$	Joint-issued primary cases	Paired DocEP/DocF1 $+.150/+.174$	External direction retained	Direct	Retain; Table S32

Claim IDs	Comparison and public margin	Audit applicability	Document evidence	Independent / temporal evidence	Final tier	Decision / source
C35 EP; C36 EF1	BioLink rerank > BioLink encoder: paired $+.041/+.040$	Same full-coverage population	Paired DocEP/DocF1 $+.059/+.057$	Direction retained across thresholds	Direct	Retain; Tables S32, S40
C37 EP; C38 EF1	BioLink encoder > PubMedBERT: paired $+.039/+.041$	Same full-coverage population	Paired DocEP/DocF1 $+.053/+.047$	Direction retained across thresholds	Direct	Retain; Tables S32, S39
C39 EP; C40 EF1	PubMedBERT > BM25+GBDT: paired $+.087/+.067$	Same full-coverage population	Paired DocEP/DocF1 $+.103/+.087$	Direction retained across thresholds	Direct	Retain; Tables S32, S38
C41 EP; C42 DocEP	Cross-rubric GPT > Claude: $+.02/+.02$	Same 600-case rubric	DocEP paired	Order retained across five seeds	Transferred	Retain; Tables S27, S28
C43 EP; C44 DocEP	Cross-rubric Claude > Qwen: $+.03/+.03$	Same 600-case rubric	DocEP paired	Order retained across five seeds	Transferred	Retain; Tables S27, S28
C45 EP; C46 DocEP	Cross-rubric Qwen > MedFound: $+.02/+.02$	Same 600-case rubric	DocEP paired	Intervals overlap	Transferred	Retain direction; widen tier
C47 EP; C48 DocEP	Cross-rubric MedFound > Llama: $+.01/+.02$	Same 600-case rubric	DocEP paired	Smallest external margin	Surrogate-only	Downgrade; Table S28

The complete-stack outcome is 36 direct, 6 transferred, and 6 surrogate-only claims, including 2 reversal flags. The 18/18 matched regime-transition claims survive at direct status; all six downgrades, including both reversals, occur in lower-margin open-row comparisons.

T DETERMINISM AND AUDIT TRACE

The combined paper and appendix enumerate the benchmark construction, audit program, compute ledger, governance protocol, and evidence traces relevant to every claim. Each main-paper statement that depends on secondary evidence points to a numbered table or figure in this appendix.

The determinism contract binds every automatic obligation to a pinned execution path. The token-level scorer, citation-budget sweeps, overlap-threshold sweeps, clean-versus-full partitioning, construct-validity re-scoring, post-cutoff re-scoring, and the 10,000-resample stratified bootstrap are deterministic given the pinned Python 3.11.7 / NumPy 1.26.3 / PyTorch 2.2.2 / FAISS 1.7.4 stack (Table S36); document-level DocEP/DocER, Likert decision-sufficiency scores, and IRB-covered clinician-panel metrics are provider-dependent and are therefore reported as frozen adjudicated ledgers rather than re-computed during review.

The audit-evidence pointer is likewise entirely internal to the PDF: the 800-output document-level DocEP/DocER audit (Table S20); the 1,600-output integrated row-stratified audit (Table S21; Figure S9); the transparent lower-bound addendum on 160 Classical hybrid outputs (Table S22); 142 adversarial citation pairs (Figure S5); the 384-case independent clinician audit (Table S37); the 612-case external slice (Table S24); the 540- and 468-case independent paths (Tables S25 and S26); the 600-case source-native cross-rubric evaluation (Table S27); the 487-case post-cutoff slice (Table S45); the 312- and 360-case cue audits (Tables S42 and S43); the open and transparent panels (Tables S39 and S40); five-seed reruns (Table S29); and the governance and source-use records (Tables S33 and S34). The matched-case trace adds 1,440 regime outputs and 3,200 row-ordering outputs, with common-population denominators, paired intervals, and raw document win/tie/loss counts in Tables S31 and S32.

U BENCHMARK-LANDSCAPE COMPARISON

Table S48 locates ADMITSCOPE against representative neighboring resource families. Medical QA contributes biomedical reasoning tasks, trial resources contribute domain-specific matching and extraction, citation benchmarks contribute attribution evaluation, model-side methods contribute evidence access and uncertainty control, and evidence-centered design contributes validity warrants. ADMITSCOPE connects these ingredients through a coverage-aware claim compiler whose output is an executed support license and source trace.

Table S48: Neighboring resource families by evaluation object, output contract, pairwise comparability, and executed claim licensing.

Resource family	Representative citations	Evaluation object	Joint output contract	Coverage-aware pairwise claim	Executed license audits
Medical QA / biomedical reasoning benchmarks	Jin et al. (2019; 2021); Gu et al. (2022); Nori et al. (2023)	Biomedical answer correctness	Dataset-dependent	Aggregate task comparison	Task robustness
Trial matching / trial NLP resources	Weng et al. (2010); Wong et al. (2023); Jin et al. (2024); Gupta et al. (2024); Marshall et al. (2020)	Matching and extraction	Dataset-dependent evidence	Task-level comparison	Domain validation
Citation-attribution benchmarks	Gao et al. (2023); Li et al. (2024)	Attribution quality	Evidence-centered output	Attribution ranking	Human and ID/OOD attribution checks
Evidence-centered benchmark design	Liu et al. (2024)	Validity-warrant framework	Design evidence	Design consideration	Warrant specification
Retrieval / calibration / selective-prediction methods	Lewis et al. (2020); Guo et al. (2017); Geifman & El-Yaniv (2017); Kompa et al. (2021)	Model component	Method-dependent	Method evaluation	Component robustness
ADMITSCOPE (this paper)	Main paper + this supplement	Three trial-operation decisions + transformed stress	Required joint output	Joint-issued and coverage-matched	Document, independent, temporal, and cross-rubric licenses