TRACEQUEUE: SOURCE-GROUNDED TRIAGE ACCELERATES PAPER–REPOSITORY VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Between repository resolution and rerun execution, an auditor must turn one queue entry into a verified follow-up decision. TRACEQUEUE makes this handoff measurable: 1,085 resolved pairs share an ordered paper-plus-repository evidence frame, a 42-field disclosure contract, calibrated risk and abstention, field-level source pointers, and a temporal rerun ledger. With GPT-4.1, evidence, temperature, calibration folds, scalar request, and thresholds held fixed, replacing a free-form rationale by the complete contract yields $\Delta\text{ROC} = -0.003$ (95% CI $[-0.018, +0.013]$), improves ECE by 0.008 ($p = 0.028$), and increases explicit abstention by 5.4 points ($p = 0.003$); 92.8% of audited Open Structured pointers identify their intended source. The decisive test is a blinded study of 126 pairs: the structured entry reduces median verification time from 17.8 to 13.6 minutes (23.6%, $p = 0.004$), lowers false follow-up from 14.6% to 8.8%, and raises verified-issue recall by 20 minutes from 64.9% to 71.6%. Paper and table evidence adds 0.055 holdout AUC over repository-only features and retains 0.048 after maturity balancing; in a fresh 204-pair cohort, full-contract risk has adjusted odds ratio 1.62 for decision-changing verified issues. Cross-team and disjoint reruns preserve the score ordering. These results show that ranking quality alone can conceal a consequential system property: how efficiently a prediction becomes a verified expert decision.

1 INTRODUCTION

Once an authoritative repository has been linked to a paper, an auditor faces a short but consequential sequence: choose what to inspect, locate the supporting paper span, table cell, or repository path, and decide whether the evidence warrants follow-up. A scalar risk score ranks the queue but does not execute this handoff. A free-form rationale can name the right concern while still forcing the auditor to search for its source. On the same evidence and GPT-4.1 backbone, that search changes the measured task from 13.6 median minutes with a structured, source-grounded entry to 17.8 with a free-form rationale, even though their holdout ROC differs by only -0.003.

This decision point sits between two better-studied tasks. PaperBench, EXP-Bench, ReproduceBench, and broad scientific-agent suites measure whether an agent can implement or execute research (Starace et al., 2025; Kon et al., 2026; Zhao et al., 2026; Bragg et al., 2026); ReproRepo measures whether static agents recover human-reported repository issues (Li et al., 2026). Document extraction, reporting frameworks, and scalar judges address other parts of the pipeline (Pineau et al., 2021; Gebru et al., 2021; Mitchell et al., 2019; Jain et al., 2020; Lo et al., 2020; 2023; Zheng et al., 2023; Tan et al., 2025). What these evaluations do not measure is the controlled conversion of identical visible evidence into a calibrated, source-located entry and then into a timed human decision. Without that control, ranking gains, extraction quality, and execution success cannot reveal whether an interface accelerates verification.

TRACEQUEUE instantiates this missing stage for 1,085 resolved pairs from NeurIPS, ICML, and ICLR. The handoff begins from a deterministic evidence frame and ends in a queue entry containing calibrated risk, a three-way decision, explicit abstention, 42 typed disclosure values, and pointers to paths, spans, or cells. A preserved codebook, temporal split, calibration procedure, and rerun identity keep the decision comparable when models or assessor teams change. The central intervention changes only the complete response contract: GPT-4.1 snapshot, evidence, temperature, scalar request, calibration folds, and thresholds remain fixed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The evaluation follows the same handoff in three steps. First, can the entry prioritize uncertainty without changing discrimination? The matched contract improves ECE by 0.008 and increases explicit abstention by 5.4 points while its ROC interval is centered near zero. Second, can an auditor resolve and act on the entry? 579/624 sampled Open Structured pointers identify the intended source, and the blinded study records a 23.6% reduction in verification time ($p = 0.004$), 5.8 points fewer false follow-ups, and a 6.7-point increase in recall by 20 minutes. Third, does the cross-artifact signal survive maturity balance and new outcomes? Paper/table fields retain a 0.048 AUC increment after balancing repository maturity, and a 204-pair outcome cohort links full-contract risk to decision-changing issues with adjusted odds ratio 1.62.

The resulting contribution is both an evaluation object and a systems finding. TRACEQUEUE supplies the resolved-pair protocol needed to compare evidence handoffs across a temporal holdout, a 312-paper cross-team rerun, and a 143-paper disjoint rerun. Its controlled studies then establish rank–handoff separation: systems with nearly identical discrimination can impose materially different verification cost. This makes source resolution, calibrated abstention, and expert decision time first-class evaluation targets rather than presentation details attached to a leaderboard score.

2 RELATED WORK

2.1 PAPER REPRODUCTION AND REPOSITORY EVALUATION

Modern reproduction benchmarks target execution capability. PaperBench grades end-to-end replication of 20 papers with hierarchical task rubrics; EXP-Bench evaluates 461 experimental tasks from 51 papers; AstaBench spans broader scientific-agent skills; and AutoReproduce evaluates generated implementations on PaperBench and the 13-paper ReproduceBench (Starace et al., 2025; Kon et al., 2026; Bragg et al., 2026; Zhao et al., 2026). ReproRepo is the closest static paper–repository task: it uses human-raised GitHub issues to evaluate whether agents surface known reproduction blockers across $1{,}149$ papers (Li et al., 2026). TRACEQUEUE asks a different, complementary question before execution: given the same resolved pair and visible evidence, which output contract produces calibrated disclosure triage that a human can verify efficiently? Its labels are rubric-based disclosure states rather than code-execution scores or issue-recovery matches, and its distinctive endpoints are pointer resolution, abstention behavior, and blinded decision time.

2.2 REPORTING AND RESEARCH-DOCUMENT RESOURCES

Computational reporting studies, checklists, and documentation frameworks characterize what papers, models, or datasets disclose (Peng, 2011; Stodden et al., 2016; Baker, 2016; Pineau et al., 2021; Dodge et al., 2019; Raff, 2019; Belz et al., 2021; Kapoor & Narayanan, 2023; Gebru et al., 2021; Mitchell et al., 2019). SciREX, S2ORC, PaperMage, and READoc instead make scientific documents machine-readable (Jain et al., 2020; Lo et al., 2020; 2023; Li et al., 2025). TRACEQUEUE composes these concerns at a decision boundary: paper spans, table cells, and repository paths populate one typed entry, and the evaluation separately adjudicates whether each emitted location resolves, identifies the intended source, supports the value, or contradicts it.

2.3 SELECTIVE PREDICTION, ATTRIBUTION, AND HUMAN–AI HANDOFF

Selective prediction and learning to defer formalize when a model should yield to an expert (Shafer & Vovk, 2008; Vovk et al., 2005; Madras et al., 2018; Mozannar & Sontag, 2020); attributed generation tests whether claims carry textual support (Gao et al., 2023). LLM-judge benchmarks expose scalar-reliability variation (Zheng et al., 2023; Fu et al., 2024; Liu et al., 2023; Tan et al., 2025), while human–AI studies show that explanations can alter reliance without improving team performance (Bansal et al., 2021). TRACEQUEUE fixes the abstention rule and compares complete handoff interfaces under a shared backbone. Direct pointer adjudication tests source binding, and the timed study measures the operational quantity that coverage, judge agreement, and explanation preference leave unobserved.

3 METHOD: THE TRACEQUEUE HANDOFF PROTOCOL

3.1 FORMAL OBJECT AND INTERVENTION

An instance is $x = (p, r, E_B)$, where p is a published paper, r its resolved authoritative repository, and E_B the paper-plus-repository evidence exposed by deterministic ordering and truncation rule B. Paper text follows reading order; tables and captions follow their first textual mention. Repository files are ordered as manifests and environment files, documented train/evaluation entry points, paths cited by the paper or README, and then remaining paths lexicographically. The repository frame stops after 128 files, and each file after 12 KB. A system returns

$$q(x) = \big(s, c, a, \mathbf{v}, \boldsymbol{\pi}\big),$$

where $s \in [0, 1]$ is calibrated risk of an incomplete disclosure, $c \in \{\text{complete}, \text{ambiguous}, \text{incomplete}\}$, $a = \mathbb{1}[s \in [0.40, 0.60]]$, $\mathbf{v} \in (\mathcal{V} \cup \{\perp\})^{42}$ contains typed field values, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{42})$ contains source locations. The fixed decision rule maps $s < 0.40$ to complete, $s > 0.60$ to incomplete, and the intervening band to ambiguous with $a = 1$. Each slot is defined by (name, value type, applicability predicate, missing-value rule, pointer type). For an applicable field, "not observed" is an explicit typed value; $\perp$ is reserved for a field whose underlying operation or claim does not exist for the paper. Both cases omit a pointer, whereas every observed value requires one. Repository pointers use `path:Lstart-Lend`, prose pointers use section plus quoted span, and table pointers use table, row, and column identifiers; all coordinates refer to E_B. Appendix A gives the field types and applicability rules.

Pointer quality is not collapsed into one label. An emitted π_j is judged for (i) resolution inside E_B, (ii) exact identification of the intended span, cell, or path, (iii) semantic support for v_j, and (iv) contradiction of v_j. A location can therefore resolve without supporting the value, and support and contradiction are adjudicated independently. This decomposition makes the handoff failure visible: the model may rank the right paper but still send the auditor to the wrong source.

The same-backbone experiment changes the complete response contract—field organization, explicit missingness, abstention, and pointers—while holding x, E_B, `gpt-4.1-2025-04-14`, temperature 0, the scalar-risk request, calibration folds, and thresholds fixed. The shared prompt prefix defines the disclosure target and presents E_B in the same order. The free-form suffix requests the scalar and a rationale; the structured suffix requests the same scalar plus the schema-conformant state, values, missingness markers, and pointers. The estimand is consequently the complete evidence-bearing interface, not a pointer-only effect. A valid rerun preserves the resolved-pair rule, B, codebook, response semantics, abstention band, and split ledger.

3.2 POPULATION AND DISCLOSURE TARGETS

The protocol follows a three-stage timeline. Stage A establishes the authoritative repository; stage B prioritizes evidence-backed follow-up; stage C performs the selected work under a team's execution protocol. TRACEQUEUE evaluates stage B, the shared point at which competing systems can receive identical visible evidence and produce comparable human-facing decisions.

The candidate frame enumerates NeurIPS, ICML, and ICLR proceedings records from 2020–2025 before label collection, without score-based sampling or venue quotas. DOI and normalized-title matches collapse duplicate proceedings records; repository identity follows an explicit paper URL, then a matching author/project URL when the paper link redirects. Conflicting repositories are retained only when adjudicators agree on one authoritative target. Table 1 records the resulting boundary: 1,085 of 1,793 candidates form authoritative clean pairs. The 994 high-completeness subset supports corruption stress tests while leaving the main target unchanged.

Three annotators rate each pair on a 1–5 reporting-completeness rubric. A $1/5$ lacks core ingredients such as executable training paths, seed policy, or environment specification; $3/5$ denotes mixed evidence with a critical omission or contradiction; $5/5$ requires jointly disclosed training recipe, environment, and ablation/runtime evidence sufficient for manual triage. Scores 2 and 4 lean toward adjacent anchors. After adjudication, scores 1–2 define the positive binary event ($y = 1$, incomplete), while scores 3–5 define $y = 0$ for ROC, PR, Brier, and ECE. The ordered target for macro-F1 and QWK keeps score 3 as the distinct ambiguous class and maps scores 4–5 to complete. Soft Brier

Table 1: Benchmark construction. The clean-pair row defines the evaluated population.

Construction stage	Removed	Remaining
Initial candidate pool (2020–2025)	—	1,793
No repository URL	432	1,361
Repository 404 / private / access-gated	107	1,254
Multiple conflicting repositories	76	1,178
Non-research content	93	1,085
Expert-rated clean-pair benchmark	—	1,085
Corruption-eligible high-completeness subset	—	994

replaces y by the fraction of the three initial raters assigning scores 1–2, retaining disagreement rather than forcing it to zero or one.

The panel contains 14 ML researchers; the reported $\kappa = 0.63$ is the mean pairwise Cohen's kappa across the three ratings per paper, and 61.7% of papers are unanimous before adjudication. A fourth senior annotator resolves every non-unanimous case without seeing model predictions. The benchmark has $411/354/320$ NeurIPS/ICML/ICLR papers, a 41.4% incomplete rate, and a strict 847-paper 2020–2024 fit/calibration versus 238-paper 2025 test split; the holdout contains $86/238$ incomplete pairs. The study received an exempt determination under 45 CFR 46.104(d)(2); annotators were co-authors or compensated above local minimum wage, could withdraw without penalty, and are reported only in aggregate.

3.3 Evidence-bearing reference implementation

The contract contains 14 repository, 12 table/figure, and 16 text-disclosure fields, operationalized in Appendix A. In redacted pair `NeurIPS-2024-r3189`, the entry cites `train.py:L41` for `torch.manual_seed(args.seed)`, `requirements.txt` for dependencies, and paper Section 4.1 for "we train on 4 NVIDIA A100 80GB GPUs." Open Structured is the non-API reference implementation. GROBID and PaperMage recover reading order, sections, captions, and cells; SciSpaCy normalizes scientific entities; the repository walker emits commit-bound paths and line spans (Lopez, 2009; Lo et al., 2023; Neumann et al., 2019). Field validators then enforce the contract type and retain explicit conflict indicators when paper and repository evidence disagree.

For scoring, each field contributes separate applicability, observation, and conflict indicators. Boolean and categorical values are one-hot encoded; counts and quantities are represented by their parsed value and unit; path and free-text values contribute typed presence rather than lexical identity. Open Structured feeds this vector to an XGBoost head, and structured GPT-4o feeds the same vector to a logistic head (Chen & Guestrin, 2016). Head selection and all preprocessing are fit on 2020–2024 only. Five venue- and label-stratified folds produce out-of-fold scores for Platt scaling; the fitted calibrator is frozen before the 2025 test and reruns (Platt, 2000). Figure 1 locates the controlled interface.

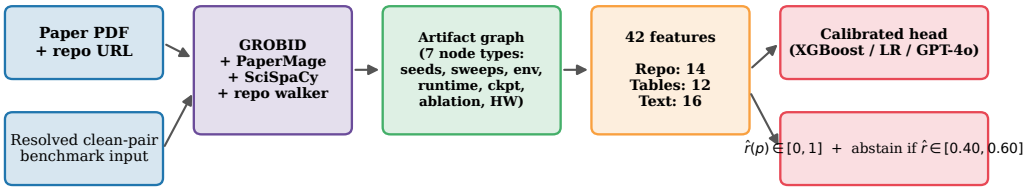

Figure 1: TraceQueue maps a resolved pair through deterministic paper/repository parsing into 42 typed fields, calibrated risk, abstention, and field-level source locations. The same evidence frame is reused across compared interfaces.

Table 2: System identity and evidentiary role. "Structured" denotes the complete field, missingness, abstention, and pointer contract.

System	Output	Controlled role	Reported in
Open Structured	42 fields + pointers, XGBoost head	Non-API reference	Main, pointer, rerun
Direct Judge	GPT-4.1 scalar + rationale	Ranking comparator	Main, outcome, rerun
GPT-4o structured	42 fields + pointers, logistic head	Strong structured comparator	Main, pointer, rerun
GPT-4.1 structured/free	Contract / rationale	Paired interface intervention	Same-backbone, utility
Full / repo-only	All fields / repository fields	Cross-artifact ablation	Matching, outcome

4 Experimental Setup

The roster separates benchmark baselines from the paired interface intervention. Open Structured is the non-API 42-field pipeline. GPT-4o fills the same contract and feeds a calibrated logistic head. Direct Judge is a benchmark-roster `gpt-4.1-2025-04-14` run that emits scalar risk plus a free-form rationale. BERT, Longformer, and Direct Judge map their calibrated scalar through the common thresholds; structured systems additionally emit the three-way state, typed fields, and pointers. The paired experiment uses separately calibrated structured and free-form GPT-4.1 arms with identical prompts up to the response-format instruction. Consequently, Direct Judge's 0.774 ROC and the paired free-form arm's 0.772 are distinct runs serving different comparisons. "Full contract" in the outcome study denotes the score using all 42 extracted fields; human-utility claims refer only to the paired GPT-4.1 arms.

For outcome validation, Checklist uses fixed rubric indicators, Paper-only uses the 16 text-disclosure fields, Repo-only uses the 14 repository fields, and Full contract uses all 42 fields with the same calibration protocol; Direct Judge supplies its scalar GPT-4.1 score. These rows score the same decision-changing endpoint. In the utility study, the repo/checklist condition presents the same deterministic E_B with the fixed checklist but without model risk, abstention, or source pointers, leaving auditors to prioritize follow-up from that static view.

All systems receive the same E_B. BERT-base uses 512 tokens; Longformer-Large-4096 uses 4,096 tokens, AdamW, learning rate 10^{-5}, batch size 4, and 3 epochs (Devlin et al., 2019; Beltagy et al., 2020). GPT-4o uses `gpt-4o-2024-05-13`. The five calibration folds preserve venue and binary-label proportions; each paper receives an out-of-fold raw score, one sigmoid calibrator is fit to those scores, and no 2025 label enters model, preprocessing, threshold, or calibration selection. We report ROC-AUC, PR-AUC, Brier score, 10-bin ECE, three-way macro-F1, and QWK (Brier, 1950; Cohen, 1968; Guo et al., 2017; Naeini et al., 2015). Intervals use 2,000 paper-stratified bootstrap resamples; stochastic heads use five seeds; paired AUC tests use DeLong's method (DeLong et al., 1988). Non-API runs use one NVIDIA A100 80GB, CUDA 12.4, PyTorch 2.3.1, Transformers 4.41.2, and Python 3.10.13. The benchmark and AUDIT-312 runs total 32.4 A100-hours, $529 for GPT-4o extraction, and $35.80 for Direct Judge; DISJOINT-143 is separately accounted.

The blinded utility study uses a venue-stratified subset of DISJOINT-143: 126 papers, 18 auditors, and 378 assignments. Every paper appears once in each of three conditions—structured GPT-4.1, its backbone-matched free-form rationale, and repo/checklist triage—with different auditors; each auditor receives seven assignments per condition and never sees the same paper twice. Assignments are blocked by venue and issue status, then randomized within auditor with balanced condition order. The evidence frame is identical across conditions. Timing starts when the entry opens and ends at the final follow-up decision; unfinished assignments are right-censored at 24 minutes. The primary statistic is the Kaplan–Meier median decision time. Crossed paper-and-auditor bootstrap resampling supplies the displayed intervals, and the structured-versus-free p value permutes condition labels within venue, issue-status, and order blocks while recomputing the censored median difference. An issue-positive paper contains an adjudicated omission or contradiction, so each condition has the

same 74-paper recall denominator. False follow-up is a completed assignment that initiates follow-up not confirmed by the adjudicated reference, divided by all completed assignments in that condition. Verified-issue recall by 20 minutes, decision accuracy, confidence calibration, and false follow-up are secondary; their intervals are descriptive, and the recall p value is not used for a separate significance claim.

Separately, 2 blinded annotators audit pointers sampled by field group from the temporal holdout and disjoint rerun. The sample is equally allocated to Open Structured and structured GPT-4o, testing the pointer semantics shared by structured implementations; the paired GPT-4.1 contract is evaluated through its timed human outcomes. The AUDIT-312 clean-pair control is the 91-paper subset with an unconflicted repository identity and all evidence paths needed by B.

Repository-maturity control treats the benchmark endpoint as the case indicator and asks whether either score discriminates after cases and controls have comparable repository profiles. Within venue-year strata, a logistic propensity model uses repository age, depth, dependency pinning, CI files, container files, checkpoints, stars, forks, and code size. Variable-ratio optimal matching retains endpoint-positive and endpoint-negative papers within a 0.20 pooled-SD caliper; unmatched papers are discarded, and retained sets must have absolute standardized mean difference at most 0.10 on every maturity covariate. Full-contract and repo-only AUC are then computed on the identical matched papers, so their difference is paired. Each of $2{,}000$ paper-stratified bootstrap replicates repeats propensity fitting, matching, and both AUC calculations. The resulting estimand is discrimination in the repository-maturity overlap population.

The outcome-linked cohort contains 204 fresh post-2025 resolved pairs; outcome adjudicators are blinded to risk score and tier. A logistic model estimates

$$\operatorname{logit} \Pr(Y_i = 1) = \alpha + \beta z_i + \gamma^\top m_i + \eta_{\mathrm{venue}(i)} + \tau_{\mathrm{year}(i)},$$

where Y_i is a decision-changing verified issue, z_i is standardized full-contract risk, and m_i contains the maturity covariates. The model includes venue and year fixed effects; missing maturity values receive a training-split median plus a missingness indicator. Paper-stratified bootstrap resampling repeats model fitting; score tertiles are fixed before outcome adjudication. The same exempt protocol, aggregate-only reporting, compensation, and withdrawal protections apply to both human studies.

5 RESULTS

5.1 AUC ALONE SELECTS A DIFFERENT HANDOFF

Table 3 reports the 2025 holdout. A leaderboard ordered only by ROC places Direct Judge (0.774 $[0.733, 0.812]$) above Open Structured (0.753); the paired difference is 0.021 (DeLong $p = 0.087$, bootstrap $[-0.005, +0.047]$). The same choice reverses on every reported decision-quality metric: Open Structured improves Brier from 0.204 to 0.193, ECE from 0.095 to 0.079, macro-F1 from 0.527 to 0.546, and QWK from 0.478 to 0.503, while attaching a source to each observed field. Moreover, 23 of the 27 papers placed in its fixed abstention band are low-agreement panel cases. The temporal test therefore motivates the controlled question that follows: whether response structure changes the handoff even when the backbone and information do not change.

5.2 THE CONTRACT CHANGES CALIBRATION WITHOUT CHANGING THE EVIDENCE

Both paired arms use `gpt-4.1-2025-04-14`, identical E_B, Platt folds, temperature, scalar request, and abstention thresholds; only the complete response contract changes. Table 4 reports structured-minus-free-form deltas. The holdout ROC estimate is -0.003 with interval $[-0.018, +0.013]$, while ECE improves by 0.008 ($p = 0.028$) and explicit abstention rises by 5.4 points ($p = 0.003$). The clean-pair rerun repeats the direction of both changes. These paired results identify calibration and selective handoff as the main interface effects of the evaluated contract, which combines organization, typed missingness, abstention, and source locations.

5.3 SOURCE-GROUNDED ENTRIES ACCELERATE VERIFIED DECISIONS

The direct audit samples 1,248 pointers from 148 papers and 4,986 emitted fields. Open Structured resolves to the intended exact source in 92.8% of cases (579/624, 90.5%–94.8%) and obtains

Table 3: Primary results on the 2025 holdout ($n = 238$). Higher is better except Brier, soft Brier, and ECE.

Method	ROC	PR	Brier	Soft Brier	ECE	F1	QWK
BERT	.681	.591	.248	.241	.113	.468	.401
Longformer	.719	.637	.214	.211	.091	.511	.459
Direct Judge	.774	.691	.204	.204	.095	.527	.478
Open Structured	.753	.670	.193	.193	.079	.546	.503
GPT-4o	.791	.718	.183	.167	.067	.583	.534

Table 4: Same-backbone contract isolation. Deltas and paired-bootstrap intervals are structured minus free form; p values use two-sided paired permutations. Lower is better for Brier and ECE.

Metric	2025 holdout			AUDIT-312 clean pair		
	Δ	95% CI	p	Δ	95% CI	p
ROC	$-.003$	[-.018,+.013]	.71	$-.004$	[-.028,+.021]	.74
Brier	$-.004$	[-.009,+.001]	.11	$-.004$	[-.012,+.004]	.32
ECE	$-.008$	[-.015,-.001]	.028	$-.008$	[-.019,+.002]	.12
Macro-F1	$+.005$	[-.008,+.019]	.43	$+.007$	[-.014,+.028]	.51
QWK	$+.007$	[-.010,+.024]	.39	$+.008$	[-.017,+.033]	.48
Abstain, pp	$+5.4$	[+2.2,+8.8]	.003	$+6.3$	[+1.5,+11.4]	.017

91.7% support precision; structured GPT-4o reaches 94.2% and 93.1%. Endpoint-level agreement is $\kappa = 0.79$ [0.74, 0.84]. Repository paths resolve most often, while table cells account for most unresolved matches (Appendix D).

Table 6 gives the human consequence. The structured condition lowers Kaplan–Meier median verification time from 17.8 to 13.6 minutes, a 4.2-minute or 23.6% saving ($p = 0.004$). By the 20-minute mark, auditors have verified issues for 53/74 issue-positive papers, compared with 48/74 under free-form rationales; the 6.7-point difference is directional ($p = 0.061$). Among completed assignments, false follow-up falls from 18/123 (14.6%) to 11/125 (8.8%). Because the intervention is the complete response contract, these effects quantify the source-grounded handoff as a whole. Figure 2 shows when the time advantage emerges.

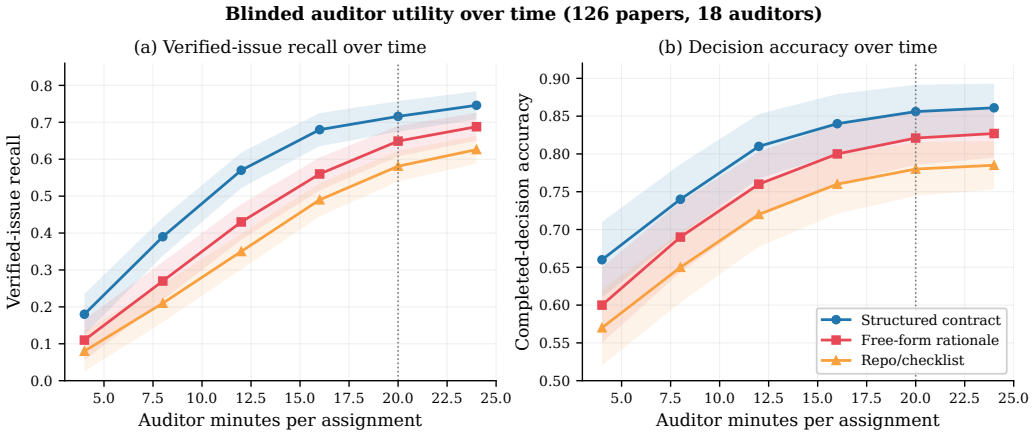

Figure 2: Verification quality versus auditor time in the blinded study. Curves aggregate counterbalanced assignments; bands are clustered 95% intervals and the vertical line marks the 20-minute recall cutoff. The complete structured contract accelerates verified decisions under the same evidence frame and backbone.

Table 5: Direct pointer audit. Exact-source resolution, support, unresolved location, and contradiction are separately adjudicated over each system's 624 sampled pointers; categories need not partition the sample.

System	Exact source	Support	Unresolved	Contradiction
Open Structured	92.8% [90.5%,94.8%]	91.7%	3.5%	1.8%
GPT-4o	94.2% [92.1%,96.0%]	93.1%	2.6%	1.4%

Table 6: Blinded auditor utility. Recall is evaluated at a 20-minute per-assignment cutoff; parentheses give paper- and auditor-clustered 95% intervals.

Condition	Recall by 20 min	Decision accuracy	Median time / assignment	False follow-up
Structured contract	71.6% (63.2–79.1)	85.6% (79.0–90.5)	13.6 min (12.2–15.1)	8.8% (4.9–15.0)
Free-form rationale	64.9% (56.5–72.8)	82.1% (74.8–87.6)	17.8 min (16.0–19.8)	14.6% (9.4–22.0)
Repo/checklist	58.1% (49.4–66.3)	78.0% (70.2–84.3)	19.4 min (17.5–21.6)	17.9% (12.0–25.7)

5.4 Paper and table evidence adds signal after maturity balancing

Open Structured exceeds repo-only by $+0.055$ AUC on the holdout $[+0.027, +0.084]$ and $+0.051$ on the 91-paper AUDIT-312 clean-pair control $[+0.019, +0.082]$. Within venue-year propensity overlaps, the full-minus-repo estimand remains $+0.048$ $[+0.022, +0.075]$, $+0.042$ $[+0.013, +0.072]$, and $+0.038$ $[+0.009, +0.068]$ on the holdout, clean-pair control, and disjoint rerun, retaining 226, 86, and 132 papers. The repeated positive differences identify paper and table evidence as incremental outcome signal after balancing the measured repository-maturity factors.

The independent outcome cohort tests the full-contract score against blinded follow-up. Of 204 pairs, 128 complete the prescribed rerun and 76 stop at an adjudicated stage; 61 expose a decision-changing verified issue. Table 7 reports discrimination and calibration for that endpoint. Full contract improves over repo-only by +0.042 AUC [+0.014, +0.073], and a one-SD increase in its score has adjusted odds ratio 1.62 [1.18, 2.24]. The prespecified tertiles are monotone: 31/68, 19/68, and 11/68 papers in descending risk order produce decision-changing findings.

5.5 The score ordering transfers across teams and time

On the 68-paper bridge, fresh auditors agree with the original panel on 56 papers ($\kappa = 0.58$, bootstrap $[0.44, 0.71]$). Open Structured remains monotone in externally authored issue count on AUDIT-312/DISJOINT-143 ($\rho = 0.50/0.49$), with top-to-bottom issue ratios $2.84\times$ / $2.63\times$ and clean-pair ratios $3.07\times$ / $2.54\times$. Table 8 shows the same method ordering in both reruns: GPT-4o has the highest AUC and rank correlation, Direct Judge is intermediate, and Open Structured carries the directly adjudicated field-level sources. These reruns establish score ordering and issue-count monotonicity across assessor groups; the paired GPT-4.1 study supplies the corresponding human-time estimate.

Table 7: Outcome-linked validation on the post-2025 cohort ($n = 204$). All rows score the same adjudicated endpoint; the maturity-adjusted full-minus-repo comparison is ΔROC $= +0.042$ [+0.014,+0.073].

Method	ROC [95% CI]	PR [95% CI]	Brier [95% CI]
Checklist	.641 [.564,.711]	.482 [.400,.565]	.232 [.204,.261]
Paper-only	.681 [.608,.748]	.531 [.449,.611]	.218 [.190,.246]
Repo-only	.706 [.635,.771]	.558 [.475,.637]	.207 [.180,.235]
Direct Judge	.735 [.669,.795]	.596 [.514,.673]	.199 [.173,.226]
Full contract	.748 [.684,.806]	.612 [.531,.687]	.191 [.166,.217]

Table 8: External reruns under the preserved pair, evidence, label, and split ledger. Intervals are 2,000 paper-stratified bootstraps.

Method	Audit AUC	Audit ρ	Disjoint AUC	Disjoint ρ
Open Structured	.721 [.665,.775]	.50	.742 [.661,.815]	.49
Direct Judge	.732 [.677,.783]	.54	.758 [.680,.826]	.52
GPT-4o	.748 [.696,.797]	.58	.773 [.698,.839]	.56

6 Discussion

The experiments expose a failure mode in ranking-only evaluation. Under matched GPT-4.1 inputs, the response contract moves calibration, abstention, and human verification time while the ROC estimate changes by -0.003 $[-0.018, +0.013]$. The 23.6% time reduction ($p = 0.004$) is the strongest operational effect, and lower false follow-up shows that speed does not come from indiscriminate escalation. Direct adjudication establishes that structured systems usually send auditors to the intended source; the outcome and maturity analyses show that the paper-plus-repository representation carries decision-relevant signal beyond repository profile. Together, these results argue for evaluating the route from score to evidence to human decision, rather than treating the interface as an interchangeable wrapper around AUC.

The evidence covers visible-disclosure triage after authoritative repository resolution across NeurIPS, ICML, and ICLR pairs. It supports decisions about which omissions or contradictions merit expert follow-up, while scientific correctness and team-specific execution outcomes use separate adjudication targets. The human study estimates the complete structured contract—organization, missingness semantics, abstention, and pointers—rather than any component in isolation. Within this regime, source-grounded structure converts the same model evidence into faster, more selective expert decisions.

7 Conclusion

TraceQueue turns the evidence handoff between repository resolution and rerun execution into a measurable systems object. On identical GPT-4.1 evidence, its structured entry reduces median verification time from 17.8 to 13.6 minutes and false follow-up from 14.6% to 8.8%, while the ROC point difference remains -0.003. The 1,085-pair temporal benchmark, direct pointer audit, maturity-balanced ablations, 204-pair outcome cohort, and two independent reruns connect that human result to calibrated, source-resolvable evidence. The broader consequence is immediate for human-facing ML evaluation: a prediction is not fully characterized by how it ranks cases, but also by how efficiently it supports a verified decision.

AI USE STATEMENT

Generative AI assistance was used in manuscript preparation, including drafting and copyediting.

REFERENCES

Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016. doi: 10.1038/533452a. URL https://doi.org/10.1038/533452a.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445717.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. doi: 10.48550/arXiv.2004.05150. URL https://arxiv.org/abs/2004.05150.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 381–393. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.29. URL https://aclanthology.org/2021.eacl-main.29/.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefanus Candra, Jason Dunkelberger, Daniel Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Ngoc-Uyen Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=M7TNf5J26u.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256. URL https://doi.org/10.1037/h0026256.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595. URL https://doi.org/10.2307/2531595.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1224. URL https://aclanthology.org/D19-1224/.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.365. URL https://aclanthology.org/2024.naacl-long.365/.

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

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

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7506–7516. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.670. URL https://aclanthology.org/2020.acl-main.670/.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu, Xinyi Zhu, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, Matei Zaharia, and Ang Chen. EXP-Bench: Can AI conduct AI research experiments? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.24785.

Shanda Li, Qiuhong Anna Wei, Jingwu Tang, Valerie Chen, Nihar B. Shah, Tim Dettmers, Yiming Yang, and Ameet Talwalkar. ReproRepo: Scaling reproducibility audits with GitHub repository issues. *arXiv preprint arXiv:2606.18237*, 2026. doi: 10.48550/arXiv.2606.18237. URL https://arxiv.org/abs/2606.18237.

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, Shanshan Jiang, Bin Dong, and Le Sun. READoc: A unified benchmark for realistic document structured extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21889–21905. Association for Computational Linguistics, 2025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The Semantic Scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447/.

Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, Amanpreet Singh, Chris Wilhelm, Angele Zamarron, Marti A. Hearst, Daniel Weld, Doug Downey, and Luca Soldaini. PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 495–507. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-demo.45. URL https://aclanthology.org/2023.emnlp-demo.45/.

Patrice Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, volume 5714 of *Lecture Notes in Computer Science*, pp. 473–474. Springer, 2009. doi: 10.1007/978-3-642-04346-8_62.

David Madras, Toniann Pitassi, and Richard S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6150–6160, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-5034. URL https://aclanthology.org/W19-5034/.

Roger D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011. doi: 10.1126/science.1213847. URL https://doi.org/10.1126/science.1213847.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

John C. Platt. Probabilities for SV machines. In *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008. URL https://doi.org/10.7551/mitpress/1113.003.0008.

Edward Raff. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL https://jmlr.org/papers/v9/shafer08a.html.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating AI's ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56843–56873. PMLR, 2025. URL https://proceedings.mlr.press/v267/starace25a.html.

Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P. A. Ioannidis, and Michela Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, 2016. doi: 10.1126/science.aah6168.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715. URL https://doi.org/10.1007/b106715.

Xuanle Zhao, Zilin Sang, Yuxuan Li, Qi Shi, Weilun Zhao, Shuo Wang, Duzhen Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. AutoReproduce: Automatic AI experiment reproduction with paper lineage. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21920–21942. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1001. URL https://aclanthology.org/2026.acl-long.1001/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A BENCHMARK AND SPLIT DETAIL

This appendix specifies the response slots and the temporal and cross-team ledgers used by every reported comparison.

Table 9: Field inventory for the 42-field response contract.

Feature group	Count	Contract fields
Repository structure	14	repository URL; resolved commit; default branch; license; train entry point; evaluation entry point; dependency manifest; version pins; environment lock; container recipe; CI workflow; seed control; checkpoint path; history depth
Tables and figures	12	data split; preprocessing; architecture; optimizer; learning-rate schedule; batch size; epochs/steps; hyperparameter search; ablation coverage; runtime; hardware/accelerators; uncertainty/error bars
Text disclosure	16	task definition; dataset identity; sample counts; split policy; metric definition; baseline identity; training objective; initialization/pretraining; stopping rule; seed policy; number of runs; compute statement; training time; inference cost; failure/exclusion criteria; evaluation protocol

Value types. Repository URL, resolved commit, default branch, and license are respectively URL, hexadecimal identifier, string, and categorical identifier. Train/evaluation entry points, dependency manifest, environment lock, container recipe, CI workflow, and checkpoint are repository-path values; version pins and seed control are typed {`observed`, `not observed`, `conflicting`} states; history depth is a nonnegative count. Data split, preprocessing, architecture, optimizer, learning-rate schedule, hyperparameter search, and ablation coverage are normalized categorical or text values. Batch size, epochs/steps, runtime, hardware, and uncertainty are quantities paired with units or aggregation semantics. Text-disclosure slots use normalized text for definitions, identities, objectives, policies, and protocols, nonnegative counts for sample and run counts, and unit-bearing quantities for compute, training time, and inference cost.

Applicability and missingness. Applicability is decided from the paper's claimed experimental operation before field extraction. Training-only slots (optimizer, schedule, batch size, epochs, seed control, checkpoint, and training time) apply when the target result requires training; inference cost applies when an inference procedure is evaluated; hyperparameter-search, ablation, and uncertainty slots apply when the corresponding empirical claim is made. Dataset, metric, split, baseline, and evaluation-protocol slots apply to every empirical comparison. A required but unseen value is `not observed`; a slot whose operation is absent is $\perp$; conflicting visible values retain the typed conflict state. Neither `not observed` nor $\perp$ may carry a pointer.

Pointer grammar and submission validity. Observed repository values use `path:Lstart-Lend`, or a bare path for whole-file evidence; prose uses `section:"quoted span"`; tables and figures use `T{id}:R{row}:C{column}` or the analogous caption span. Coordinates are resolved only against E_B, so external URLs cannot satisfy a pointer. A structured submission is malformed if it omits any slot, emits a value outside its type, attaches a pointer to a missing value, or omits a pointer from an observed value. Malformed entries receive no field-level credit but retain their scalar score for discrimination evaluation.

Prediction and calibration map. Open Structured encodes applicability, observation, value, and conflict components as described in Section 3; GPT-4o and structured GPT-4.1 populate the identical contract. Direct Judge, BERT, and Longformer emit a raw scalar. Every raw scalar is calibrated from the same five 2020–2024 folds, and the frozen $0.40/0.60$ thresholds map it to complete, ambiguous/abstain, or incomplete. Binary metrics use adjudicated scores 1–2 as positive and 3–5 as negative; ordered metrics preserve score 3 as ambiguous. These rules make output mappings comparable without requiring every baseline to emit fields.

Table 10: Benchmark and external rerun splits. Benchmark rows report incomplete rate; rerun rows report papers with at least one externally authored issue.

Split	Papers	Rate	Overlap	Use
Train (2020–2024)	847	42.9%	—	Fit/calibrate
Test (2025)	238	36.1%	None	Primary test
AUDIT-312	312	226/312 (72.4%)	68-paper bridge	Cross-team rerun
DISJOINT-143	143	91/143 (63.6%)	None	Disjoint rerun

Table 11: Main benchmark composition. Agreement is unanimous versus adjudicated.

Venue	Train	Test	Total	Incomplete rate	Agreement
NeurIPS	312	99	411	39.2%	261/150
ICML	278	76	354	43.5%	217/137
ICLR	257	63	320	41.9%	191/129
Total	847	238	1,085	41.4%	669/416

B FULL SAME-BACKBONE RESULTS

Table 12: Complete structured versus free-form GPT-4.1 comparison. Both arms share backbone, retrieval, calibration, and abstention thresholds.

Split	Arm	ROC	PR	Brier	ECE	F1	QWK	Abstain
2025	Structured	.769	.686	.198	.083	.534	.493	10.1%
2025	Free-form	.772	.689	.202	.091	.529	.486	4.7%
AUDIT-312 clean	Structured	.744	.668	.207	.089	.518	.468	12.4%
AUDIT-312 clean	Free-form	.748	.674	.211	.097	.511	.460	6.1%

Table 12 exposes the operating points behind the paired deltas. Across both splits, the structured arm improves Brier, ECE, macro-F1, and QWK and routes roughly twice as many cases through explicit abstention, while ROC changes by 0.003–0.004. The distributions below place those differences against paper-matched uncertainty.

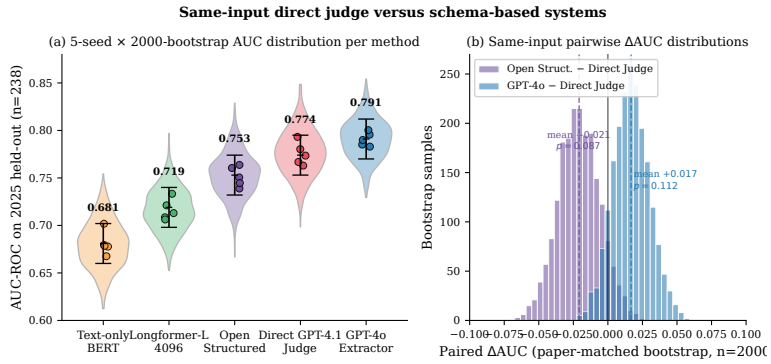

Figure 3: Same-input uncertainty on the 2025 holdout. Five-seed, paper-stratified bootstrap distributions place Direct Judge at 0.774 AUC and Open Structured at 0.753 and quantify each paired contrast.

The distribution view complements Table 3: point estimates determine the displayed ordering, while paired resampling shows the uncertainty attached to each contrast and prevents a small ranking gap from being mistaken for the handoff effect.

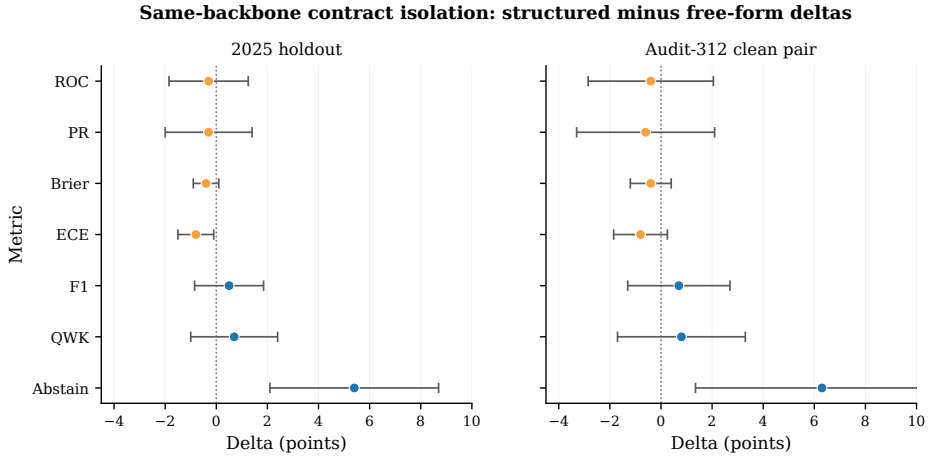

Figure 4: Structured-minus-free-form same-backbone deltas. Calibration and abstention exhibit the clearest changes; the paired discrimination estimates remain near zero.

Across the holdout and clean-pair rerun, calibration and explicit deferral move in the same direction while the discrimination intervals remain centered near zero.

C Blinded Auditor Study Detail

Each of the 74 issue-positive papers contributes one independently assigned decision to each condition. By 20 elapsed minutes, the structured, free-form, and repo/checklist conditions verify 53/74, 48/74, and 43/74 of those papers. Across all 378 assignments, completed decisions are 125/126, 123/126, and 123/126; incomplete decisions remain censored at 24 minutes for Kaplan–Meier time analysis and are excluded from completed-decision accuracy. Accuracy numerators are 107/125, 101/123, and 96/123; false-follow-up numerators are 11/125, 18/123, and 22/123. The multiway bootstrap independently resamples paper and auditor clusters and recomputes condition medians; the blocked permutation keeps each observed time/censoring pair intact while exchanging interface labels. The displayed kappa compares each condition's decisions with the adjudicated reference, not two auditors assigned to the same paper-condition cell.

Table 13: Additional human-utility endpoints with clustered 95% intervals. Time is in minutes.

Endpoint	Structured	Free-form	Repo/checklist
Time to first verified issue	6.8 [5.9,7.8]	9.7 [8.4,11.2]	11.4 [9.9,13.1]
Confidence ECE	.071 [.049,.097]	.102 [.076,.132]	.128 [.097,.163]
Decision–reference κ	.72 [.64,.80]	.61 [.52,.70]	.55 [.45,.64]

D Pointer Resolution Detail

2 blinded annotators independently judge whether each pointer resolves, whether it identifies the intended exact source, whether that source supports the emitted value, and whether it contradicts the value; a senior annotator adjudicates disagreements. Every rate uses the system's full 624-pointer stratified sample as denominator. Open Structured identifies the exact source for 579/624 pointers; structured GPT-4o does so for 588/624, or 94.2% [92.1%,96.0%]. Support precision is 91.7% [89.0%,93.9%] for Open Structured and 93.1% [90.8%,95.1%] for GPT-4o. Unresolved rates are 3.5% [2.2%,5.4%] and 2.6% [1.5%,4.1%]; contradiction rates are 1.8% [0.9%,3.2%] and 1.4% [0.6%,2.7%]. Support and contradiction are semantic judgments and may overlap location-error categories.

Table 14: Exact-source resolution by field group, with paper-stratified 95% intervals.

Field group	Open Structured	Structured GPT-4o
Repository	96.1% [93.4,97.9]	96.8% [94.2,98.3]
Tables/figures	90.4% [86.8,93.4]	92.1% [88.7,94.8]
Text disclosure	91.8% [88.2,94.6]	93.6% [90.4,96.0]

E Outcome and External Uncertainty Detail

For the maturity control, incomplete pairs are cases on the benchmark splits and papers with a decision-changing issue are cases in the outcome cohort. The propensity model and matching use no system score. Matching is rerun inside each bootstrap, and both full and repo-only AUCs use the same matched weights; the reported $+0.048/+0.042/+0.038$ contrasts are therefore within-overlap paired differences. Retained sample sizes are 226, 86, and 132.

The staged outcome ledger records 187/204 build completions, 151/204 training completions, and 128/204 evaluation completions. Adjudicated failures occur at build for 17, training for 36, and evaluation for 23 pairs. The cohort contains three equal risk tiers of 68, 68, and 68 papers. Decision-changing issue rates are 45.6% [34.2,57.3], 27.9% [18.7,39.6], and 16.2% [9.3,26.9]. Repository-maturity strata contain 68, 71, and 65 papers, with 25, 22, and 14 decision-changing issues.

Paired AUC differences are +.011 [-.024,+.047] and +.027 [-.011,+.066] for Direct Judge and GPT-4o minus Open Structured on AUDIT-312. The corresponding DISJOINT-143 differences are +.016 [-.030,+.061] and +.031 [-.016,+.079]. These paired intervals quantify ranking separation while Table 8 shows stable ordering across teams.

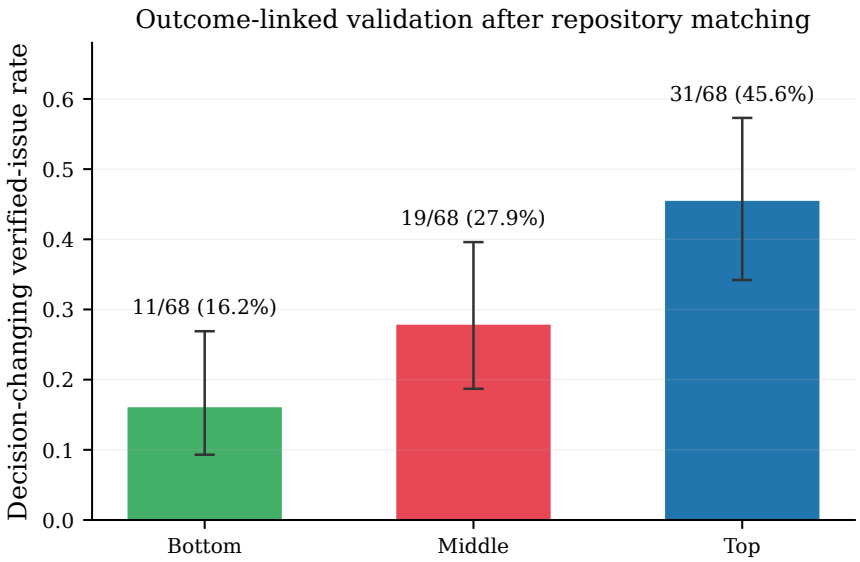

Figure 5: Decision-changing verified issues by TRACEQUEUE risk tier in the post-2025 outcome cohort. Intervals are paper-stratified; all three tiers contain 68 papers.

F DECISION-INTERFACE EVIDENCE CHAIN

Table 15: Questions, matched comparisons, and decision consequences.

Question	Comparison	Finding
Does structure alter the score?	Same GPT-4.1, E_B, calibration, and thresholds; contract versus rationale	ROC -0.003, ECE -0.008, abstention $+5.4$ points (Table 4)
Do pointers reach evidence?	Fixed pointer semantics; Open Structured and structured GPT-4o	92.8%/94.2% exact-source resolution (Tables 5, 14)
Does handoff help auditors?	Same evidence/backbone; three blinded conditions	23.6% lower time, 5.8 points fewer false follow-ups (Table 6)
Do paper fields add signal?	Maturity-balanced full versus repo-only scores	AUC gains $+0.048/+0.042/+0.038$ (Section 5)
Does risk predict outcomes?	Blinded outcome adjudication; five scoring systems	+0.042 AUC, 1.62 adjusted odds ratio (Table 7)
Does ordering transfer?	Preserved ledger with fresh assessors	$\rho = 0.50/0.49$ across reruns (Table 8)

The comparisons deliberately answer different links in the handoff. The paired GPT-4.1 row identifies the complete response interface; pointer adjudication measures source-location behavior in structured systems; maturity, outcome, and rerun analyses test whether the underlying signal survives beyond the original temporal split. No result is transferred across systems or endpoints.