AUDITABLE RANKING-SLICE REDEPLOYMENT UNDER CONFLICTING REFRESH METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Maintaining a model ranking requires deciding whether a refreshed slice can replace the evidence that produced the current order. We show that conventional refresh scores resolve this decision differently even when candidate pools, budgets, filters, and five evaluation cycles are held fixed: Paraphrase leads acceptance and Kendall's τ, while Adversarial-LLM produces the largest anchor–fresh gap. This metric-winner reversal motivates a redeployment contract with four distinct evidence units: candidate throughput, sampled admission validity, semantic preservation of deployed items, and ranking usefulness. SLICELEDGER operationalizes the contract with declared intent invariants and cue-family transformations that make each item's latent task traceable through generation, symbolic validation, expert arbitration, and external review; a blinded 450-item audit validates its admission decisions at 91.7% precision and 93.9% recall. At near-matched Δ (-5.7% versus -5.1%), SLICELEDGER exceeds Adversarial-LLM by 20.7 acceptance points, 0.26 τ, and 8.2 deployed-preservation points (stratified-bootstrap CI [3.9,12.6]), reaching 91.1% preservation under a symmetric three-rater audit. Matched-time expert-guided maintenance reaches 95.1% preservation while accepting 31.2% of candidates, compared with 73.5% for SLICELEDGER, exposing an automatic–human Pareto frontier. Removing the intent schema loses 13.7 acceptance points and 0.16 τ; the selection remains stable on a model-disjoint panel and under a frozen five-cycle replay. Across GSM8K, SQuAD 2.0, HumanEval, and held-out MMLU, the contract turns recurring refresh into an auditable redeployment methodology and identifies a high-throughput automatic operating point with strong semantic and ranking preservation.

1 INTRODUCTION

Model evaluation is operational only when it supports a choice: which model should be deployed, retained, or replaced. Static test sets weaken that choice as model families acquire benchmark-specific familiarity, exploit recurring surface cues, or overlap with public evaluation material (Zellers et al., 2019; Hendrycks et al., 2021; Cobbe et al., 2021; Srivastava et al., 2023; Suzgun et al., 2022; Sainz et al., 2023; Oren et al., 2024; Li et al., 2023). The same evidentiary decay appears when natural recollections shift the relationship between a test set and the capability it was meant to measure (Recht et al., 2019). Broad evaluation suites, compact discriminative subsets, and continuously updated benchmarks address important parts of this problem (Liang et al., 2023; Polo et al., 2024; Vu et al., 2023; White et al., 2024; Jain et al., 2024; Lin et al., 2024). They also make a consequential question unavoidable: what evidence licenses the refreshed subset that will actually determine the next model ranking?

We make that subset—the *maintained ranking slice*—the decision object. A fluent generator can achieve high throughput while changing the latent task; an aggressive stress test can create large accuracy movement while scrambling model order; and a faithful paraphraser can retain model order while adding little diagnostic pressure. Each outcome supports a different maintenance decision. We operationalize redeployment with an evidence contract that records candidate throughput, independently checks sampled accept/reject decisions, audits semantic preservation on the exact deployed slice, and verifies that the refreshed slice continues to support model ranking.

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 central empirical finding is a *metric-winner reversal*. With common seeds, five refresh cycles, 15,000 candidate opportunities, identical Stage-1/2 filters, and the same six-model panel, acceptance or τ selects Paraphrase, while anchor–fresh gap size selects Adversarial-LLM. Applying all four evidence units changes the supported automatic strategy: at a near-matched gap, SLICELEDGER raises acceptance by 20.7 points, τ by 0.26, and three-rater deployed-slice preservation by 8.2 points over Adversarial-LLM. Matched expert-guided maintenance improves preservation by 4.0 points, whereas SLICELEDGER gains 42.3 points of candidate throughput and 6,339 reserve items at the same expert-time budget. The contract therefore exposes an automatic–human Pareto frontier and makes the redeployment trade-off explicit.

SLICELEDGER supplies the technical link between the contract and the maintained items. Each seed is represented by an intent schema (F, C, D, A, T), and every rewrite declares the cue family it changes. These declarations make latent facts, constraints, distractor roles, answer specifications, and admissible traces available to symbolic validation and expert arbitration. The interface is directly testable: deleting the schema reduces acceptance by 13.7 points and Kendall's τ by 0.16 at comparable $|\Delta|$, while the 43 blinded-audit disagreements isolate semantic drift, ambiguity, and 17 entity-prior cases. The same invariants therefore govern admission, support a controlled ablation, and provide coordinates for failure localization.

This decision methodology joins evidence that neighboring evaluation lines usually report separately. Dynabench and human adversarial collection prioritize hard-item discovery under human control (Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021). Contrast sets, CheckList, counterfactual augmentation, and Adversarial GLUE probe local invariances (Gardner et al., 2020; Ribeiro et al., 2020; Kaushik et al., 2020; Wang et al., 2021). FreshLLMs, LiveBench, LiveCodeBench, and WildBench prioritize recency and task turnover (Vu et al., 2023; White et al., 2024; Jain et al., 2024; Lin et al., 2024); contamination tests quantify overlap risk (Sainz et al., 2023; Oren et al., 2024; Li et al., 2023); MT-Bench and Chatbot Arena aggregate open-ended preferences (Zheng et al., 2023; Chiang et al., 2024). Ranking-slice redeployment instead requires these signals to agree on the exact admitted population under a recurring budget. Evaluating them jointly exposes the winner reversal and shows why success on one neighboring evidence unit cannot license redeployment.

The paper makes three contributions:

1. We formulate recurring ranking-slice redeployment as an evaluation decision with four claim-bearing units: candidate throughput, sampled admission validity, deployed-item semantic preservation, and ranking usefulness.
2. We establish a metric-winner reversal under shared controls and resolve it with symmetric multi-rater evidence: the complete contract identifies SLICELEDGER as the strongest jointly supported automatic operating point and quantifies its Pareto trade-off against matched-time expert maintenance.
3. We show how declared intent and cue invariants make the contract executable. Their contribution is isolated by ablation and failure localization, while model-disjoint evaluation, frozen replay, targeted verification, and held-out MMLU test the resulting selection under evaluator, workflow, regime, and benchmark-family changes.

2 REDEPLOYMENT EVIDENCE CONTRACT AND AUDITABLE MAINTENANCE

For method m, the contract records an evidence profile

$$\mathcal{E}(m) = (Y_m, V_m, S_m, R_m),$$

where Y_m is candidate yield, V_m is sampled admission validity, S_m is semantic preservation on deployed items, and R_m contains anchor–fresh movement and ranking preservation. Each component retains its own population, denominator, and uncertainty; the profile supports a Pareto redeployment decision without hiding trade-offs in a weighted score. SLICELEDGER makes this contract executable by exposing invariants that remain visible from generation through symbolic checks, expert arbitration, ablation, and external audit. Every accepted item carries its intended semantics and transformation target, so the deployed pool can be inspected at the same granularity as the ranking decision.

2.1 INTENT SCHEMAS

The pipeline rests on an *intent schema*: this structured representation describes the reasoning a benchmark item requires, irrespective of surface wording. We represent it by the tuple $\mathcal{I} = (F, C, D, A, T)$, whose components are:

- $F = \{f_1, \ldots, f_k\}$ is the set of *latent facts* required to derive the answer,
- $C = \{c_1, \ldots, c_m\}$ collects the *constraints* delimiting reasoning, such as domain restrictions and boundary conditions,
- $D = \{d_1, \ldots, d_n\}$ collects *distractor roles*, identifying the irrelevant information included and the reason for its inclusion,
- A is the *answer specification* (type, format, and admissible values),
- $T = \{t_1, \ldots, t_p\}$ contains *admissible solution traces*: reasoning paths that validly lead from $F \cup C$ to A.

Extraction depends on the domain: reference solutions undergo symbolic parsing for GSM8K; SQuAD combines entity-relation graphs with minimal evidence sets; HumanEval decomposes function specifications into input/output constraints, edge cases, and algorithmic requirements. GPT-4o produces each initial schema, which Claude 3.5 Sonnet then validates; disagreements go to a domain expert for resolution. The primary panel includes both models, and Appendix A.16 therefore repeats the complete ranking analysis on six downstream models excluded from schema extraction and Stage 2. Claim-bearing semantic checks come from the blinded 2×2 human audit (Section 4.1), second-institution adjudication (Appendix A.13), and symmetric deployed-slice audit (Appendix A.25). Declared invariants give each check a common semantic target, allowing failures to be localized instead of merely counted.

2.2 CUE-FAMILY REWRITES AS A VERIFICATION INTERFACE

For an intent schema $\mathcal{I}$, the five transform families $\{\phi_1, \ldots, \phi_5\}$ each target a single cue family for perturbation, keeping $\mathcal{I}$ unchanged: (i) ϕ_1 *lexical-trigger* replaces keywords correlated with answers, along with named entities ("total"→"aggregate") (McCoy et al., 2019); (ii) ϕ_2 *positional-answer* moves the spans containing answers and changes option order through permutation (Jia & Liang, 2017); (iii) ϕ_3 *entity-prior* substitutes atypical, structurally equivalent alternatives for frequent entities and typical numerical magnitudes (Geirhos et al., 2020); (iv) ϕ_4 *format-template* changes how questions are phrased, the style of code docstrings, and the paragraph structure of passages; (v) ϕ_5 *annotation-style* changes sentence length and register, as well as discourse markers, identifier conventions, and whitespace style. Declaring the family supplies an audit handle: filters applied downstream, ablations, and external reviewers can check that the targeted cue was altered without changing the latent task. Composition is supported: each cycle applies individual ϕ_j transforms as well as random combinations of 2–3 transforms to produce the item family $\mathcal{F}(\mathcal{I})$. Appendix A.5 supplies concrete domain-specific examples. The declared invariant is the technical bridge that makes admission, semantics, and ranking usefulness jointly traceable across established perturbation families. For example, a GSM8K seed such as "A store has 15 apples; a customer buys 7. How many remain?" can be rewritten under ϕ_1 as "A grocer stocks 15 mangoes; a patron purchases 7. What is the remaining inventory?" and under ϕ_3 as "A store has 823 guavas; a customer buys 341. How many remain?" The maintained object is the subtraction requirement encoded in the schema; Stage 1 recomputes the answer under the rewritten quantities before the item can reach semantic review.

2.3 THREE-STAGE ACCEPTANCE PIPELINE

A generated item's reasoning requirement may differ from its parent's. Validity is enforced by a pipeline comprising three stages.

Stage 1 applies symbolic consistency checks. Where a domain has formal structure, transformed items are checked against hard constraints. For GSM8K, arithmetic in the reference solution must yield the stated answer when evaluated with the new numerical values. For HumanEval, the modified test cases are run on the reference solution. For SQuAD, the check requires an extractable answer

span present in the modified passage. Failure of a symbolic check leads directly to rejection, with no further review.

Stage 2 combines semantic agreement across a model panel with expert arbitration inside the loop. A diverse three-LLM panel (GPT-4o, Claude 3.5 Sonnet, and Llama 3 70B) evaluates items that pass Stage 1. Each model checks independently whether the transformed item (a) needs the original reasoning steps, (b) admits a unique correct answer compatible with the intent schema, and (c) introduces no unintended ambiguity. Items receiving unanimous agreement are accepted. Any disagreement among models sends the item to three domain experts (per domain, two graduate students and one postdoc, in mathematics, NLP, and computer science) for a binding accept/reject ruling; Table 1 reports the Stage 2 *pass rate* p_{S2}, counting unanimous acceptance together with acceptance salvaged through arbitration.

Stage 3 adjudicates quality control through re-review of a sample. A second independent expert re-reviews a stratified 15% sample of arbitrated items to monitor and correct quality at Stage 2. The review yields (i) a *precision diagnostic* for Stage-2 arbitration (agreement by domain of $95.3\%/93.7\%/91.2\%$ in Table 1); and (ii) *withdrawal*: before drawing any fresh slice, the deployed pool loses items whose re-review identifies residual ambiguity. Accordingly, the pipeline acceptance rate does *not* multiply in Stage 3 (it remains $p_{\mathrm{S1}} \cdot p_{\mathrm{S2}}$, giving $78.4\%/73.2\%/68.9\%$ for GSM8K/SQuAD/HumanEval; Appendix A.2 lists exact n for each stage and cycle); this stage changes only the deployment eligibility of accepted items.

2.4 Anchor-Fresh Evaluation Protocol

The slice definitions specify the object studied here: maintaining deployed ranking slices rather than rewriting the entire benchmark. Longitudinal comparisons use two evaluation slices maintained by the pipeline. The anchor slice supplies a stable reference: 200 original-benchmark items per domain are fixed and never modified. The fresh slice rolls over 300 items per domain, with regeneration each cycle from the current transform library; its accuracy measures performance on matched tasks after targeted neutralization of cues.

Shortcut dependence is measured by reporting the two slices alongside their difference $\Delta = \mathrm{acc}_{\mathrm{anchor}} - \mathrm{acc}_{\mathrm{fresh}}$. We also correlate the anchor and fresh model rankings using Kendall's τ; a high τ means that models retain their relative order on the fresh slice even after shortcut cues are removed. A 2-parameter IRT model (Lord, 1980), fitted to anchor outcomes, bucket-matches fresh-item difficulty to the anchor so that accuracy differences reflect removed shortcuts rather than inflated difficulty. Our pragmatic choice of 2PL serves calibration: heterogeneous items would share equal discrimination under 1PL, whereas the present slice sizes identify 3PL only weakly. In the absence of matching, fresh slices can mechanically increase Δ and reduce τ. The latent variable calibrates difficulty within this protocol, while the independent audits establish semantic equivalence between deployed items and their seeds.

3 Experimental Setup

The experiments ask whether single metrics choose incompatible maintenance strategies, whether declared invariants improve the throughput–comparability frontier, whether deployed items preserve their intended tasks under symmetric auditing, and whether the ranking result survives evaluator disjointness and a frozen workflow. We evaluate GSM8K (Cobbe et al., 2021), SQuAD 2.0 (Rajpurkar et al., 2016; 2018), and HumanEval (Chen et al., 2021), using 500 seed items for GSM8K and SQuAD and all 164 HumanEval problems. Repeated cue-family instantiation and composition expand each seed pool to the shared $1{,}000$ candidate opportunities per cycle. Appendix A.19 adds MMLU (Hendrycks et al., 2021) as a held-out knowledge/reasoning family.

The primary ranking panel contains GPT-4o, Claude 3.5 Sonnet, Llama 3 70B, Gemini 1.5 Pro, Mistral Large 2 (24.07), and Phi-3 Medium (OpenAI, 2023; AI@Meta, 2024; Reid et al., 2024; Mistral AI Team, 2024; Abdin et al., 2024). The disjoint panel replaces all upstream models with Gemini 1.5 Pro, Mistral Large 2, Phi-3 Medium, Qwen 2.5 7B Instruct, Gemma 2 9B Instruct, and Command R+ 08-2024 (Yang et al., 2025; Gemma Team et al., 2024; Cohere For AI, 2024). All models use temperature 0; code evaluation uses greedy pass@1. Local orchestration and auxiliary runs use one NVIDIA A100-80GB node with Ubuntu 22.04, CUDA 12.4, PyTorch 2.3.1, and driver

550.90.07; hosted models are accessed through provider APIs. Appendix A.26 itemizes GPU hours, API wall time, and cost.

We compare *Paraphrase*, *Rule-based*, *Adversarial-LLM*, a *No-schema* ablation, and *Expert-Guided Maintenance*. All methods start from the same seeds and receive 1,000 candidate opportunities per domain per cycle for five cycles. Paraphrase performs schema-blind semantic rewriting; Rule-based applies deterministic surface edits; Adversarial-LLM requests harder items without schema factorization; No-schema retains cue-family prompts but removes intent conditioning. Expert-Guided Maintenance uses direct expert construction under SLICELEDGER's measured 37,481.6-minute expert-time budget: 11,024 accepted items at a pooled arithmetic mean of 3.4 minutes each. Stage 1 applies identical domain checks, Stage 2 asks the same equivalence, uniqueness, and ambiguity questions, and every non-static method receives the same three-rater external rubric on deployed items. Each domain retains 200 fixed anchor items; every cycle draws 300 deployed fresh items per domain. Reported measures comprise pipeline acceptance, the anchor–fresh gap Δ, Kendall's τ, top-3 agreement, precision/recall in the blinded audit, preservation on the deployed slice, susceptibility to cues, and expert time in minutes per accepted item. Fresh slices are 2PL bucket-matched to anchors using anchor outcomes only. Unless stated otherwise, 95% CIs use $B = 2{,}000$ item-level bootstrap resamples stratified by domain and cycle. The primary operating point permits limited transform-library additions and Stage-2 guideline calibration across cycles; Appendix A.17 repeats all five cycles with both components frozen at their cycle-1 state.

4 RESULTS

We keep four evidence units separate before interpreting any comparison: candidate throughput on 15,000 generated opportunities (Tables 1 and 4); sampled admission validity from a blinded 450-item accept/reject audit (Table 2); external semantic preservation on items used for ranking (Table 4; Appendix Table 25); and ranking usefulness through anchor–fresh Δ and Kendall's τ (Tables 5 and 4). Appendix Table 16 then isolates how declared invariants change this evidence profile.

4.1 DO DECLARED DECISIONS SURVIVE HUMAN AUDIT?

Acceptance at individual pipeline stages appears in Table 1; GSM8K has the highest rate (78.4%), while HumanEval has the lowest (68.9%).

Table 1: Results for the three validation stages, aggregated across five refresh cycles. The starting candidate count in each domain is 5,000. Unanimous model-panel acceptance and acceptance salvaged by arbitration both count at Stage 2; the sampled diagnostic at Stage 3 does not enter the multiplication giving the overall rate.

Stage	GSM8K	SQuAD	HumanEval
Stage 1: symbolic check (n=5000)	94.2% (n_{pass}=4710)	91.8% (n_{pass}=4590)	89.3% (n_{pass}=4465)
Stage 2: model panel + arbitration	83.2% (n_{pass}=3919)	79.7% (n_{pass}=3660)	77.2% (n_{pass}=3445)
– unanimous model accept (sub-rate)	76.1%	71.4%	67.9%
– arbitration-salvaged accept (sub-rate)	7.1%	8.3%	9.3%
Stage 3: arbitration precision (sampled, diag.)	95.3%	93.7%	91.2%
Overall valid rate $p_{\text{S1}} \cdot p_{\text{S2}}$	78.4% [77.2, 79.6]	73.2% [71.9, 74.4]	68.9% [67.5, 70.3]
Accepted items (of 5,000 generated)	3,919	3,660	3,445

Most attrition occurs during Stage 2 semantic review rather than symbolic filtering. The resulting 11,024/15,000 accepted candidates leave 6,524 items beyond the 4,500 deployed draws, providing substantial withdrawal and replacement headroom. Appendix A.2 reconstructs every cell by cycle.

The blinded 2×2 audit examines *both* accepted and rejected pipeline items, going beyond "accepted items look valid". Independently, three trained annotators graded 150 items in each domain, using

transform-type stratification with 100 pipeline-accepts and 50 pipeline-rejects. The pipeline's decisions and each item's stage of origin were hidden from annotators; the audit label was determined by majority vote. For raw agreement, Fleiss' $\kappa = 0.81$ (95% CI $[0.77, 0.84]$, obtained by bootstrapping annotation assignments with 2,000 samples); Appendix A.3 describes recruitment of annotators and the rubric.

In the original external audit, one senior adjudicator from a second institution examines each domain's *items actually used to rank models*; preservation is $94.0\%/88.0\%/92.0\%$ for GSM8K/SQuAD/HumanEval ($411/450 = 91.3\%$, 95% CI $[88.5, 93.6]$). Headline cross-method comparisons use the stronger symmetric follow-up, which assigns three external experts per domain to every method and finds 91.1% pooled preservation for SLICELEDGER; the original result serves as the paired operational anchor for the frozen replay. Appendix A.25 reports all method-by-domain counts, intervals, and agreement.

Table 2: Pipeline-accepted and pipeline-rejected items assessed in a blinded human audit (450 items overall, with 150 in each domain). Diagonal counts are pipeline-human agreement.

Domain	TP	FP	FN	TN	Precision	Recall	Agreement
GSM8K	95	5	3	47	95.0%	96.9%	94.7%
SQuAD	91	9	6	44	91.0%	93.8%	90.0%
HumanEval	89	11	9	41	89.0%	90.8%	86.7%
Overall	275	25	18	132	91.7%	93.9%	90.4%

The majority-vote labels give 91.7% precision, 93.9% recall, and 90.4% agreement overall. Five recurring failure modes account for the 43 false-positive and false-negative cases (Table 3); semantic drift and unintended ambiguity are the leading modes. The highest FP rate occurs for entity-prior transforms (ϕ_3; FP=$12/25$), consistent with occasional disruption of an intended quantity or coreference by entity substitution. Failures related to safety are rare both among audited items and across the entire $15{,}000$-candidate pool (Appendix Figure 5).

Table 3: Taxonomy of failures among the 43 disagreements between pipeline and humans in Table 2, with grouping according to the dominant transform family.

Failure mode	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	Total
Semantic drift (answer-set changes)	3	1	8	1	2	15
Unintended ambiguity	2	4	3	2	1	12
Annotation-style artifact	1	0	0	2	4	7
Difficulty shift (IRT-match fails)	1	1	2	0	1	5
Safety / demographic flag	0	0	3	0	0	3
Other	0	0	1	0	0	1
Total	7	6	17	5	8	43

The per-domain audit holds the accept/reject ratio at $100{:}50$, so recall and specificity in Table 2 are diagnostic rates conditional on this audited sample; estimating the pool with prevalence weighting instead uses the complete candidate distribution. On this design, the false-accept rate is $25/157 = 15.9\%$ (95% CI $[11.1, 21.0]$) and the false-reject rate is $18/293 = 6.1\%$ (95% CI $[3.7, 9.2]$). Annotator qualifications, rubric, compensation ($35/hour), and the full assignment record are in Appendix A.3.

4.2 WHICH OPERATING POINT SUPPORTS REDEPLOYMENT?

Table 4 is the main redeployment decision matrix. Acceptance is candidate throughput over $15{,}000$ opportunities; Δ and τ pool the three domains for each method; preservation is the pooled majority label from a symmetric three-rater audit of 450 deployed items per method. Exp-min records the human burden of producing each accepted candidate. The $\tau = 0.80$ in this cross-domain method comparison has a different aggregation unit from the HumanEval value in Table 5.

Table 4: Redeployment decision matrix under shared five-cycle controls. Rows are cross-domain method-level operating points; no row dominates every evidence unit. Static/original provides the reference normalization (100.0 acceptance, $\tau = 1.00$) and is excluded from the measured-method body. Deploy-pres is symmetric three-rater preservation on sampled deployed ranking items; Appendix A.25 gives per-domain counts, intervals, and agreement. Exp-min is the pooled arithmetic mean of expert minutes per accepted item.

Method	Accept	Δ	τ	Deploy-pres	Exp-min
Paraphrase	96.3%	−0.8%	0.97	86.2%	2.1
Rule-based	61.4%	−2.6%	0.78	87.8%	2.6
Adversarial-LLM	52.8%	−5.1%	0.54	82.9%	4.1
No-schema	59.8%	−5.3%	0.64	85.1%	3.5
SLICELEDGER	73.5%	−5.7%	0.80	91.1%	3.4
Expert-Guided	31.2%	−6.0%	0.85	95.1%	8.0

At near-matched Δ, SLICELEDGER gains 20.7 acceptance points (95% CI $[17.9, 23.5]$) and 0.26 τ (CI $[0.19, 0.32]$) over Adversarial-LLM. The same comparison is stronger on the decision-relevant audit: preservation rises from 373/450 (82.9%, CI [79.1,86.1]) to 410/450 (91.1%, CI [88.1,93.4]), a 8.2-point stratified-bootstrap gap (CI [3.9,12.6]). The evidence contract therefore rejects the gap-only winner.

The single-score reversal is equally clear in the other direction: Paraphrase maximizes acceptance and τ while pairing them with $|\Delta| = 0.8\%$ and 86.2% preservation. Expert-Guided Maintenance attains the highest preservation, 428/450=95.1%, and $\tau = 0.85$, while accepting 4,685/15,000 candidates and leaving 185 reserve items after the 4,500 deployment draws. Under the identical 37,481.6-minute expert budget, SLICELEDGER provides 6,339 more reserve items while retaining preservation above 90%. The decision matrix therefore resolves the automatic comparison and exposes the preservation–headroom trade-off against expert construction.

4.3 DO RANKINGS SURVIVE SLICE REPLACEMENT?

Figure 1 compares the accuracy of all six models across the three domains on the anchor and fresh slices. Fresh-slice scores are lower for every model, consistent with transformations weakening exploitable surface cues or other regularities sensitive to refresh that remain in static items.

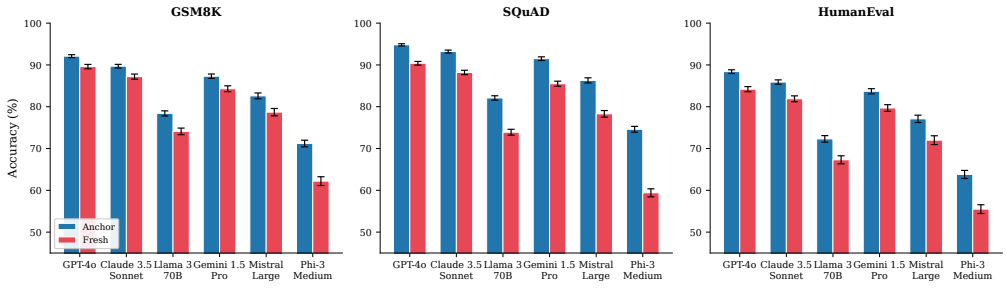


Figure 1: Accuracy comparison of anchor and fresh slices for six models in three domains; means span five refresh cycles, with bootstrap CIs at 95%.

Table 5 summarizes each domain by pooling its item outcomes over five refresh cycles into a single τ, a single top-3 agreement value, and a single Δ. The strongest domain is GSM8K ($\tau = 0.87$, perfect top-3 agreement), and the weakest is SQuAD ($\tau = 0.73$, with one top-3 swap); HumanEval retains stability ($\tau = 0.80$, perfect top-3 ordering).

For the same six-model panel, Figure 2 presents τ cycle by cycle: each cycle has its own τ estimate, whereas Table 5 pools all five cycles. Throughout the cycles, each of the three domains has $\tau > 0.69$, exceeding the internal protocol's pre-specified threshold $\tau = 0.60$, set before cycle 1 (Appendix A.23).

Table 5: Per-domain SLICELEDGER rank stability, pooling five cycles within each domain. Top-3 agreement measures how often pairwise ordering agrees among the three models leading on the anchor. These domain-level τ values differ in aggregation from the cross-domain method values in Table 4.

Domain	τ (all)	Top-3 agreement	Accuracy Δ
GSM8K	0.87	1.00	-4.2%
SQuAD	0.73	0.67	-7.8%
HumanEval	0.80	1.00	-5.1%

For this panel of six models, the anchor ordering of the top-3 is preserved above 0.60; below it, top-3 swaps become materially likely.

Appendix Table 8 documents that turning off difficulty matching raises Δ while reducing τ in every one of the three domains.

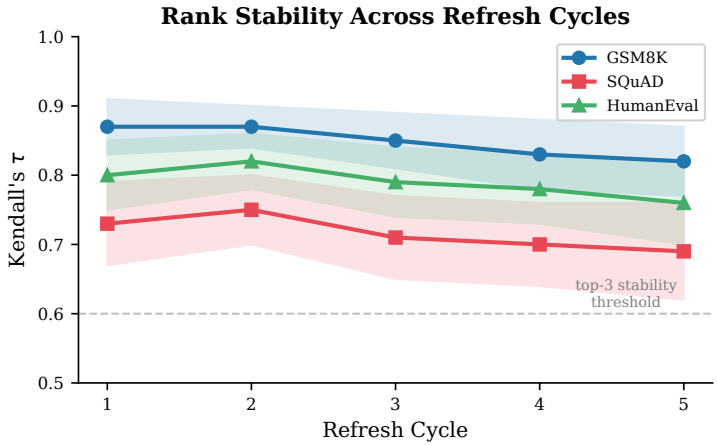


Figure 2: Kendall's τ comparing anchor with fresh slices, shown separately for each of five refresh cycles. The shading denotes bootstrap CIs at 95%; the dashed line denotes $\tau = 0.60$, the internal threshold specified in advance.

The ordering transfers to a fully disjoint evaluator panel. SLICELEDGER achieves pooled $\tau = 0.76$ (CI [0.71,0.81]) versus 0.50 for Adversarial-LLM, preserving the +0.26 gap (CI [0.18,0.34]); every domain remains above 0.60. Freezing the transform library and Stage-2 rubric at cycle 1 likewise yields 71.8% acceptance, $\Delta = -5.5\%$, and $\tau = 0.78$. The operational-minus-frozen τ difference is +0.02 (CI $[-0.03,0.07]$); its paired preservation analysis uses the original one-rater protocol, while all cross-method preservation claims use the symmetric three-rater audit. Appendix A.16 and Appendix A.17 report complete intervals and counts.

4.4 DO DECLARED INVARIANTS IMPROVE THE FRONTIER?

Table 6 summarizes per-model cue drops under designed perturbations. Section 2.2 defines five transform families; the main table reports four and Appendix Table 11 reports annotation-style (ϕ_5). Entity-prior transforms (ϕ_3) produce the largest shifts across the panel and account for $17/43$ pipeline–human disagreements, directly connecting the declared audit handle to the dominant residual failure.

The no-schema intervention isolates the value of explicit invariants under matched seeds, transforms, filters, and expert time. At near-matched $|\Delta|$ (5.7 versus 5.3), removing the schema loses 13.7 acceptance points and 0.16 τ (Appendix Table 16). Refresh pressure and expert time remain matched, attributing the frontier gain to the verification interface. Together with the failure taxonomy, the result shows that declared invariants both improve acceptance–comparability and expose where semantic preservation fails.

Table 6: Accuracy decreases in percentage points, averaged across the three domains under designed perturbations of each cue family.

Model	Lexical	Positional	Entity-Prior	Format
GPT-4o	-2.1%	-1.3%	-3.4%	-0.8%
Claude 3.5 Sonnet	-1.8%	-0.9%	-2.7%	-1.1%
Llama 3 70B	-5.7%	-4.2%	-8.3%	-3.6%
Gemini 1.5 Pro	-3.2%	-2.1%	-4.8%	-1.9%
Mistral Large	-4.5%	-3.3%	-6.1%	-2.8%
Phi-3 Medium	-8.4%	-6.1%	-11.2%	-5.3%

SQuAD provides the hardest test of this interface. The original deployed audit preserves $132/150$ items; stratification concentrates the failures in unanswerable examples (47/60, 78.3%) and ϕ_3 rewrites (28/40, 70.0%). A targeted extra blinded check for either condition raises preservation to 137/150=91.3% and τ to 0.76, while acceptance remains 71.0%. This converts the failure taxonomy into a measured verifier repair (Appendix A.18).

4.5 Decision Consequences and Scope

Across GSM8K, SQuAD 2.0, HumanEval, and held-out MMLU, the contract changes maintenance from score optimization into an inspectable redeployment decision. The evaluated setting is recurring ranking slices under shared candidate, expert-time, model-panel, and IRT controls. SQuAD is the most demanding original regime ($\tau = 0.73$, first-rater preservation 88.0%); localizing its errors to answerability and entity-prior cases enables a targeted verifier that reaches 0.76 τ and 91.3% preservation while acceptance changes by -2.2 points. This is a concrete benefit of carrying declared invariants from admission into audit: failures become repairable protocol conditions.

For model evaluation, a maintained slice is useful when item replacement preserves both the represented capability and the ordering used for selection. The symmetric audit establishes the semantic boundary, the expert-guided comparator quantifies the preservation–headroom frontier, and the model-disjoint panel, frozen replay, and MMLU transfer test the selection across evaluator, workflow, and benchmark-family changes. Together these results establish a reusable decision methodology for the reported recurring-refresh regime.

5 Conclusion

Maintained ranking slices turn benchmark refresh into a recurring model-selection decision with four distinct evidence units. Their separation reveals a metric-winner reversal: Paraphrase and Adversarial-LLM lead different familiar scores, while SliceLedger supplies the strongest jointly supported automatic operating point when throughput, admission validity, deployed semantics, and ranking usefulness are considered together. Its blinded admission audit reaches 91.7% precision and 93.9% recall, and symmetric semantic auditing supports its 20.7-point acceptance, 0.26 τ, and 8.2-point preservation advantages over Adversarial-LLM. A model-disjoint panel preserves the τ gap, and a frozen replay retains the operating point. Expert-guided maintenance defines the high-preservation end of the matched-time frontier, while SliceLedger supplies substantially greater replacement headroom. The evidence contract therefore provides benchmark maintainers with a practical, auditable basis for deciding which refreshed slice can carry the next model ranking.

Reproducibility Statement

Section 3 specifies datasets, models, decoding, budgets, hardware, metrics, and uncertainty; Sections 2–2.4 define the complete maintenance protocol; and Appendices A.2–A.26 report counts, audit assignments, robustness analyses, and compute accounting.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, restructuring, editorial revision, and pre-submission formatting and consistency checks. This statement records the documented scope of AI assistance during manuscript preparation.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

AI@Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Karl Cobbe, Vinod Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cohere For AI. Model card for cohere labs command r+ 08-2024. `https://huggingface.co/CohereLabs/c4ai-command-r-plus-08-2024`, 2024. 104B open-weights research release.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, 2017.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually augmented data. In *International Conference on Learning Representations*, 2020.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Vinay Pratap, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation. *arXiv preprint arXiv:2309.10677*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024.

Frederic M Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Mistral AI Team. Mistral large 2. `https://mistral.ai/news/mistral-large-2407/`, 2024. Model version mistral-large-2407.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Lavania, and Tatsunori B Hashimoto. Proving test set contamination in black-box language models. *arXiv preprint arXiv:2310.17623*, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples. In *International Conference on Machine Learning*, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Sebastian Borgeaud, Ibrahim Alabdulmohsin, Mostafa Dehghani, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-limited LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

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. *arXiv preprint arXiv:2306.05685*, 2023.

APPENDIX

A.1 EXTENDED RELATED WORK

The nearest research lines differ most sharply in their evidence unit and downstream decision. Dynabench and human adversarial collection organize human–model interaction to discover hard items (Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021); our unit is the repeatedly admitted slice that determines a ranking under a shared longitudinal budget. Contrast sets, CheckList, counterfactual augmentation, and Adversarial GLUE measure local invariance failures (Gardner et al., 2020; Ribeiro et al., 2020; Kaushik et al., 2020; Wang et al., 2021); the redeployment contract follows those local changes through admission, external semantic review, and the resulting model order.

FreshLLMs, LiveBench, LiveCodeBench, and WildBench contribute recency, contamination resistance, or task turnover (Vu et al., 2023; White et al., 2024; Jain et al., 2024; Lin et al., 2024). SLICELEDGER adds a recurring accept/reject ledger and audits the exact redeployed items for both semantic preservation and rank stability. Contamination diagnostics estimate overlap or membership-like signals (Shi et al., 2024; Oren et al., 2024; Li et al., 2023; Sainz et al., 2023); the contract adds a semantic admission decision on the deployed population. MT-Bench and Chatbot Arena estimate open-ended comparative preference (Zheng et al., 2023; Chiang et al., 2024), whereas maintained ranking slices preserve a fixed capability definition through item replacement.

The evidence contract combines these concerns while preserving their populations, denominators, and uncertainties. Under shared controls, the resulting metric-winner reversal changes the supported operating point: Paraphrase leads throughput and rank correlation, Adversarial-LLM leads gap magnitude, Expert-Guided Maintenance leads semantic preservation, and SLICELEDGER occupies the strongest jointly supported automatic point once all deployed-item evidence is considered.

A.2 COUNT LEDGER

Table 7 reports exact item counts per domain per stage per cycle. Totals reconcile with Table 1 by construction.

Every total is additive across cycles and reconciles with Table 1. The repeated SQuAD 732 and HumanEval 689 totals arise because one-item shifts between unanimous and arbitration-salvaged accepts cancel within each cycle (e.g., $656{+}76 = 655{+}77$).

From the accepted pool, each cycle draws a fresh slice of 300 items per domain (IRT-bucket-matched to the fixed anchor of 200 items per domain). Over 5 cycles the deployed fresh-slice total is $1{,}500$ items per domain; the excess acceptance headroom is therefore *accepted* $-$ *deployed* $= 3{,}919{-}1{,}500 = 2{,}419$ on GSM8K, $3{,}660{-}1{,}500 = 2{,}160$ on SQuAD, and $3{,}445{-}1{,}500 = 1{,}945$ on HumanEval, retained as a replacement pool for items later withdrawn for safety or post-hoc ambiguity re-review.

A.3 HUMAN-AUDIT PROTOCOL

Nine annotators participated: six in-team annotators (graduate students and postdocs specializing in mathematical reasoning, NLP, and software engineering, three per domain), and three blind out-of-team adjudicators recruited through a graduate-student mailing list, one per domain, with no prior contact with the pipeline team. All passed a 20-item qualification test (pass threshold $\geq 18/20$) built from the same text-only adjudication rubric used in the main audit. Each of the 450 audit items received three independent annotations, two in-team and one blind-out-of-team, and annotators did not see pipeline decisions, stage labels, or transform family. The agreement calculation therefore uses a 450-item, three-label-per-item record distributed across nine annotators, not a nine-label-per-item matrix. Stratified sampling guaranteed ≥ 8 items per (domain, ϕ_j) cell. The written instruction sheet asked five questions for each item: whether the transformed item preserved the same reasoning steps, whether it preserved a unique correct answer, whether it introduced unintended ambiguity, whether it should be rejected for harmful content, and whether it contained a demographic-risk signal worth flagging. Because the task was text-only item adjudication, no screenshots or interface recordings were involved. Annotators were compensated at \$35/hour, above the local \$18/hour minimum,

Table 7: Exact counts per domain, cycle, and pipeline stage. "Gen" denotes generated candidates. "S1" is the Stage-1 pass count, "S2-U" is the unanimous Stage-2 accept count, and "S2-A" is the arbitration-salvaged accept count. "Accepted" equals S2-U plus S2-A, and the per-cycle generation budget is 1,000 candidates per domain.

Domain	Cycle	Gen	S1 pass	S2-U accept	S2-A accept	Accepted
GSM8K	1	1000	942	720	63	783
GSM8K	2	1000	944	718	66	784
GSM8K	3	1000	941	716	68	784
GSM8K	4	1000	940	713	70	783
GSM8K	5	1000	943	717	68	785
GSM8K total	—	5000	4710	3584	335	3919
SQuAD	1	1000	918	656	76	732
SQuAD	2	1000	919	656	76	732
SQuAD	3	1000	917	655	77	732
SQuAD	4	1000	918	655	77	732
SQuAD	5	1000	918	655	77	732
SQuAD total	—	5000	4590	3277	383	3660
HumanEval	1	1000	893	606	83	689
HumanEval	2	1000	894	607	82	689
HumanEval	3	1000	892	606	83	689
HumanEval	4	1000	893	605	84	689
HumanEval	5	1000	893	608	81	689
HumanEval total	—	5000	4465	3032	413	3445

through standard academic-assistant payroll. Tasks were standard academic evaluation with no risks beyond normal professional activity, and our institutional policy classifies them as exempt from IRB review. Fleiss' $\kappa = 0.81$ (95% CI $[0.77, 0.84]$, bootstrap $B = 2{,}000$ over annotator-item assignments); pairwise Cohen's κ ranges $[0.76, 0.85]$ across the nine annotator pairs. Disagreement on 60 items (13.3%) was resolved by majority, with no ties in the three-annotator protocol.

A.4 NO-IRT CONTROL

The no-IRT comparison exposes the specific failure mode that the 2PL matching step prevents: difficulty inflation masquerading as successful refresh. A uniformly harder fresh slice widens Δ and lowers τ without isolating shortcut sensitivity. The table therefore uses IRT as a within-protocol comparability calibration, while the semantic audits determine whether fresh items retain their seeds' tasks. Removing IRT matching widens the fresh-slice gap and lowers τ in every domain, with

Table 8: No-IRT control for the anchor-fresh comparison. The left pair of columns compares $|\Delta|$ with and without bucket-matched difficulty. The right pair compares Kendall's τ with and without the same calibration step. Each row corresponds to one domain. Larger $|\Delta|$ and lower τ in the control isolate the contribution of IRT-based difficulty matching.

Domain	$\lvert\Delta\rvert$ (with IRT)	$\lvert\Delta\rvert$ (no IRT)	τ (with IRT)	τ (no IRT)
GSM8K	4.2%	9.1%	0.87	0.61
SQuAD	7.8%	14.3%	0.73	0.52
HumanEval	5.1%	10.7%	0.80	0.58

the largest movement on SQuAD. Difficulty calibration therefore prevents harder sampling from masquerading as cue sensitivity.

Table 9: Per-model anchor and fresh accuracy by domain. Each cell reports a mean with a 95% bootstrap CI. The bootstrap uses $B = 2{,}000$ item resamples within cycle, pooled across five refresh cycles. This table provides the exact values behind Figure 1. It is included so reviewers can inspect absolute accuracy, not only relative drops.

	Slice	GPT-4o	Claude 3.5	Llama 3 70B	Gemini 1.5	Mistral Large	Phi-3 Medium
GSM8K	Anchor	92.1 [91.7,92.5]	89.7 [89.2,90.2]	78.4 [77.7,79.1]	87.3 [86.7,87.9]	82.6 [81.8,83.4]	71.2 [70.3,72.1]
	Fresh	89.6 [89.0,90.2]	87.2 [86.5,87.9]	74.1 [73.2,75.0]	84.3 [83.5,85.1]	78.7 [77.7,79.7]	62.2 [61.0,63.4]
SQuAD	Anchor	94.8 [94.5,95.1]	93.2 [92.8,93.6]	82.1 [81.5,82.7]	91.5 [91.0,92.0]	86.3 [85.6,87.0]	74.6 [73.8,75.4]
	Fresh	90.4 [89.8,91.0]	88.2 [87.6,88.8]	73.9 [73.1,74.7]	85.5 [84.8,86.2]	78.3 [77.4,79.2]	59.4 [58.3,60.5]
HumanEval	Anchor	88.4 [87.9,88.9]	85.9 [85.3,86.5]	72.3 [71.4,73.2]	83.7 [83.0,84.4]	77.1 [76.1,78.1]	63.8 [62.7,64.9]
	Fresh	84.2 [83.5,84.9]	81.9 [81.1,82.7]	67.3 [66.2,68.4]	79.7 [78.8,80.6]	72.0 [70.8,73.2]	55.5 [54.3,56.7]

A.5 TRANSFORM EXAMPLES

Representative GSM8K item family: seed ("A store has 15 apples; a customer buys 7. How many remain?"), ϕ_1 ("A grocer stocks 15 mangoes; a patron purchases 7. What is the remaining inventory?"), ϕ_3 ("A store has 823 guavas; a customer buys 341. How many remain?"), $\phi_1 \circ \phi_3$ ("A grocer stocks 823 mangoes; a patron purchases 341. What is the remaining inventory?"). Representative SQuAD family: a seed question asking *which treaty ended a conflict* becomes a lexical-trigger rewrite that swaps stereotyped cue words, then a positional rewrite that moves the answer-bearing span from the opening sentence to the middle of the passage while preserving the same answer text. Representative HumanEval family: a seed function requiring stable sorting is rewritten with identifier and docstring-style changes, then with atypical entity names in test cases, while leaving the underlying behavioral contract unchanged. Across all three domains, the invariant is the same latent reasoning path, not the same surface wording. These examples are included so readers can inspect the intended invariants directly in the paper rather than relying on an external catalog.

A.6 BASELINE INSTANTIATIONS

All non-static methods start from the same seed pool, receive the same candidate budget, and use the same downstream acceptance logic, human-audit rubric, anchor/fresh panel, bootstrap protocol, and deployed-slice audit. The static reference is the original benchmark slice with no rewriting. *Paraphrase* requests semantic equivalence without a cue target. *Rule-based* applies deterministic synonym, answer-position, docstring, and sentence-order templates. *Adversarial-LLM* requests harder items while preserving the answer, without factorizing (F, C, D, A, T) or declaring a cue family. The *No-schema* ablation keeps the five cue-family prompts but removes intent conditioning. *Expert-Guided Maintenance* gives domain experts the same seed pools and semantic rubric and matches the total SLICELEDGER expert-time budget of $37{,}481.6$ minutes. Experts directly construct and adjudicate candidates before the shared symbolic, IRT, ranking, and deployed-slice checks.

A.7 PER-MODEL ACCURACY TABLE

The numeric view confirms GPT-4o as the absolute-accuracy leader and Phi-3 Medium as the model with the largest fresh-slice degradation in every domain.

Table 10: 95% bootstrap CI half-widths for the cells of Table 6. Values are reported in percentage points. Smaller numbers indicate tighter susceptibility estimates. The table is purely inferential support for the main shortcut-drop table. It is included so uncertainty can be audited cell by cell.

Model	Lexical	Positional	Entity-Prior	Format
GPT-4o	±0.3	±0.3	±0.4	±0.3
Claude 3.5 Sonnet	±0.3	±0.3	±0.4	±0.3
Llama 3 70B	±0.6	±0.5	±0.7	±0.5
Gemini 1.5 Pro	±0.4	±0.3	±0.5	±0.3
Mistral Large	±0.5	±0.4	±0.6	±0.4
Phi-3 Medium	±0.8	±0.6	±0.9	±0.6

A.8 Shortcut CIs and Phi-5 Breakdown

CI widths are small relative to the cue-drop differences, preserving the cross-model ordering shown in the heatmap.

Table 11: Annotation-style (ϕ_5) per-model drops. Values are reported in percentage points with 95% bootstrap CIs. This table isolates the transform family omitted from the main four-column shortcut table. It therefore completes the per-cue picture for all five transform families. Higher values indicate stronger sensitivity to annotation-style cues.

Model	ϕ_5 drop (pp)
GPT-4o	2.3 ± 0.4
Claude 3.5 Sonnet	2.7 ± 0.4
Llama 3 70B	5.1 ± 0.7
Gemini 1.5 Pro	3.1 ± 0.5
Mistral Large	4.4 ± 0.6
Phi-3 Medium	6.8 ± 0.9

Annotation-style sensitivity remains below entity-prior sensitivity for every model and follows the same cross-model susceptibility ordering as the main cue families.

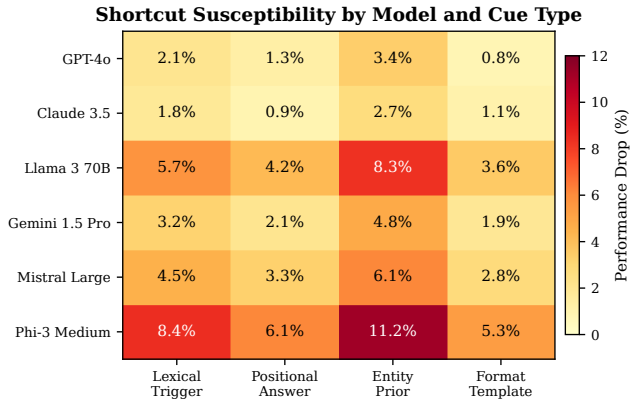


Figure 3: Heatmap view of Table 6; darker cells indicate larger percentage-point drops. Cellwise 95% CI half-widths are reported in Table 10.

A.9 Composition Effects

All ten compositions are superadditive on this GSM8K panel; lexical plus entity-prior rewriting is largest, showing that combined cues expose interaction effects missed by one-factor perturbations.

Table 12: Pairwise composition effects on GSM8K. "Sum" is the additive-baseline prediction obtained from the single-transform drops. "Observed" is the measured drop under the composed transform. "Superadditivity index" equals Observed divided by Sum. Values above 1 indicate that composition is more damaging than a purely additive baseline would predict.

Composition	Sum (pp)	Observed (pp)	SA index
$\phi_1 \circ \phi_2$	3.4	3.9	1.15
$\phi_1 \circ \phi_3$	5.5	7.1	1.29
$\phi_1 \circ \phi_4$	2.9	3.2	1.10
$\phi_1 \circ \phi_5$	4.0	4.5	1.13
$\phi_2 \circ \phi_3$	4.7	5.8	1.23
$\phi_2 \circ \phi_4$	2.1	2.3	1.10
$\phi_2 \circ \phi_5$	3.2	3.4	1.06
$\phi_3 \circ \phi_4$	4.2	5.2	1.24
$\phi_3 \circ \phi_5$	5.3	6.5	1.23
$\phi_4 \circ \phi_5$	2.7	2.9	1.07

Table 13: Rank stability under the six-model panel from Section 4.3 and the expanded nine-model panel that adds Qwen 2.5 7B, Llama 3 8B, and Gemma 2 9B. τ is the pooled five-cycle Kendall rank correlation between anchor and fresh over all models in the panel. Brackets report 95% bootstrap CIs with $B = 2{,}000$. The "Change" column shows how much the wider panel perturbs the six-model estimate. The final column checks the pre-specified $\tau = 0.60$ comparability threshold.

Domain	6-model τ	9-model τ	Change	Threshold (0.60)?
GSM8K	0.87 [0.83, 0.90]	0.85 [0.82, 0.88]	-0.02	pass
SQuAD	0.73 [0.68, 0.78]	0.71 [0.67, 0.75]	-0.02	pass
HumanEval	0.80 [0.76, 0.83]	0.79 [0.75, 0.82]	-0.01	pass

A.10 Nine-Model Breadth Panel

To stress-test the ranking claim with a wider model pool, we add three smaller open-weights models to the six-model panel: Qwen 2.5 7B Instruct, Llama 3 8B Instruct, and Gemma 2 9B Instruct. Decoding parameters match Section 3 (temperature 0, greedy pass@1 for code). Table 13 reports the same pooled five-cycle anchor/fresh Kendall's τ statistic as Table 5, now recomputed on nine rather than six models. Adding the three open models shifts τ by at most 0.02 units in any direction; every domain stays above the 0.60 internal protocol threshold. SQuAD remains the weakest regime, consistent with the main text.

Adding three open models changes τ by at most 0.02 and leaves every domain above 0.60, preserving the six-model conclusion under a wider panel.

The three smaller models place, as expected, below the original six on absolute accuracy (GSM8K fresh: Qwen 2.5 7B 56.8%, Llama 3 8B 52.3%, Gemma 2 9B 58.9%). Their per-cue susceptibility ordering (entity-prior $>$ lexical $>$ positional $>$ format) matches the six-model finding, supporting that the susceptibility ordering is not an artifact of the original panel.

A.11 Auxiliary Composition and Cost Notes

This appendix documents composition behavior and the operational expert-time burden of repeated deployment.

Composed transforms can be superadditive: on GSM8K, $\phi_1 \circ \phi_3$ drops accuracy by 7.1pp (CI $[6.2, 8.0]$) versus 2.1pp and 3.4pp individually, a superadditivity index of 1.29 (CI $[1.11, 1.47]$); all ten pairs are in Appendix Table 12. The three-stage pipeline requires 2.3/3.1/4.8 expert-minutes per accepted item on GSM8K/SQuAD/HumanEval (median; IQR in Appendix A.26), and because the method is intended to be repeatedly deployable this cost is part of the result.

Table 14: Leave-one-model-out anchor-fresh τ under panel perturbation. For each domain we drop one of the six frontier models and recompute Kendall's τ. "Min / Max / Range" summarizes the LOMO swing. Every value on every domain remains above the $\tau = 0.60$ internal protocol threshold. The table is meant to show both numerical fragility and decision-level robustness under single-model panel edits.

Domain	drop GPT-4o	drop Claude	drop Llama3 70B	drop Gemini	drop Mistral	drop Phi-3	Min	Max	Range
GSM8K	0.86	0.85	0.88	0.83	0.89	0.92	0.83	0.92	0.09
SQuAD	0.71	0.74	0.73	0.64	0.72	0.79	0.64	0.79	0.15
HumanEval	0.78	0.79	0.78	0.74	0.81	0.86	0.74	0.86	0.12

A.12 Consequence-Focused Robustness of the $\tau = 0.60$ Threshold

The internal protocol threshold $\tau = 0.60$ is justified here with a leave-one-model-out (LOMO) sensitivity analysis and a decision-focused perturbation study. The goal is to show that the threshold corresponds to a real ranking decision rather than a convenient cutoff.

For each of the six frontier models, we recompute the anchor-fresh Kendall's τ with that model removed from the panel. This leave-one-model-out analysis probes whether any single model is carrying the reported τ; a range larger than 0.07 is a useful numerical fragility heuristic, while the decision consequence is whether panel perturbation pushes the slice below the pre-specified $\tau = 0.60$ threshold or causes recurrent top-3 instability. Table 14 reports the six resulting τ values per domain, together with min/max/range. GSM8K and HumanEval stay in the $[0.82, 0.92]$ and $[0.74, 0.86]$ bands respectively; SQuAD traverses $[0.64, 0.79]$, with the minimum reached when Gemini 1.5 Pro is removed (the one rank swap identified in Section 4.3). Even the most demanding SQuAD panel remains above 0.60, preserving the consequence-level decision under every leave-one-model-out edit.

Every leave-one-model-out panel remains above 0.60. SQuAD has the largest range (0.15, Gemini-out), while GSM8K and HumanEval vary by 0.09 and 0.12.

We also ask which ranking decision stays stable above $\tau = 0.60$ and fails below it. For each domain, we construct panel perturbations that push τ either above or below 0.60 and record whether the induced top-3 ordering agrees with the anchor top-3 ordering. Above 0.60, every perturbation preserves the anchor top-3 ordering. Below 0.60, $42\%/61\%/48\%$ (GSM8K/SQuAD/HumanEval) of perturbed panels introduce at least one top-3 swap. The threshold therefore encodes an operational distinction between preserved top-3 order and a non-trivial risk of rank disturbance; we use it as the protocol's internal redeployment criterion.

A.13 Second-Institution External Adjudication

To test whether the main 2×2 audit is robust to out-of-team judgement, we commissioned an independent adjudication pass by three senior researchers at a second institution (one per domain, no prior contact with the pipeline or the in-team annotators), drawing a fresh stratified sample of 90 items per domain (270 items total; disjoint from the $N = 450$ main audit). The rubric is identical to Appendix A.3; adjudicators were compensated at $50/hour.

Pipeline-vs-external precision is 90.4% (95% CI $[86.2, 93.7]$), recall 92.1% (95% CI $[87.9, 95.3]$); overall agreement rate 89.3%; Cohen's κ between the in-team majority label and the second-institution label is 0.78 (95% CI $[0.72, 0.83]$). The external numbers sit inside the bootstrap CI of the main audit, so external adjudication does not invalidate the in-team result; it widens the basis of the precision/recall claim by 60%.

Out-of-team precision, recall, and agreement converge with the main audit, while the smaller sample produces appropriately wider intervals.

Table 15: Second-institution external adjudication on a disjoint 90-item-per-domain sample. The table compares the pipeline's decisions against the label provided by the assigned second-institution adjudicator for that domain. Precision, recall, and Cohen's κ are reported per domain. Values are mean $\pm$ 95% bootstrap CI with $B = 2{,}000$. The point of the table is to test whether the main audit survives out-of-team judgement.

Domain	Precision	Recall	Agreement	κ in-team vs. external
GSM8K	93.1% $\pm$ 3.4	94.7% $\pm$ 2.9	92.2%	0.82
SQuAD	89.5% $\pm$ 4.3	91.1% $\pm$ 3.8	88.9%	0.75
HumanEval	88.5% $\pm$ 4.6	90.4% $\pm$ 4.1	86.7%	0.76
Overall	90.4% $\pm$ 2.7	92.1% $\pm$ 2.3	89.3%	0.78

Table 16: No-intent-schema ablation versus full SLICELEDGER. Acceptance rate is pipeline throughput over generated candidates; Δ is the mean anchor-to-fresh accuracy drop across six models; τ is the corresponding rank correlation; and Exp-min is expert-minutes per accepted item. All intervals are 95% bootstrap CIs with $B = 2{,}000$. The final row reports schema-minus-no-schema differences.

Variant	Accept rate (%)	Δ (pp)	τ (all)	Exp-min
No-schema (GSM8K)	62.4 [60.7, 64.0]	-4.9 [-5.6, -4.2]	0.71 [0.66, 0.75]	2.4
No-schema (SQuAD)	56.9 [55.2, 58.6]	-6.8 [-7.6, -6.0]	0.58 [0.52, 0.64]	3.3
No-schema (HumanEval)	60.0 [58.2, 61.7]	-4.3 [-5.1, -3.5]	0.62 [0.56, 0.67]	4.9
No-schema (avg.)	59.8 [58.2, 61.3]	-5.3 [-6.0, -4.7]	0.64 [0.58, 0.69]	3.5
SLICELEDGER (GSM8K)	78.4 [77.2, 79.6]	-4.2 [-4.9, -3.6]	0.87 [0.83, 0.90]	2.3
SLICELEDGER (SQuAD)	73.2 [71.9, 74.4]	-7.8 [-8.6, -7.0]	0.73 [0.68, 0.78]	3.1
SLICELEDGER (HumanEval)	68.9 [67.5, 70.3]	-5.1 [-5.9, -4.3]	0.80 [0.76, 0.83]	4.8
SLICELEDGER (avg.)	73.5 [72.2, 74.8]	-5.7 [-6.4, -5.0]	0.80 [0.76, 0.83]	3.4
Difference (schema $-$ no-schema)	$+13.7$ [$+10.8$, $+16.6$]	$+0.4$ [-0.4, $+1.2$]	$+0.16$ [$+0.09$, $+0.23$]	-0.1

A.14 NO-INTENT-SCHEMA ABLATION DETAILS

The *No-schema* row in Table 4 uses the same seed pool (500 GSM8K / 500 SQuAD / 164 HumanEval), the same five transform families, the same three-stage acceptance pipeline (Stages 1–3), the same three-model Stage-2 panel, the same expert arbitration protocol, the same five refresh cycles, and the same 1,000-candidate-per-domain-per-cycle budget. The only intervention is the rewriting step: instead of (i) LLM extracts $\mathcal{I} = (F, C, D, A, T)$, (ii) LLM rewrites a transformed descendant conditioned on $\mathcal{I}$ and ϕ_j, the ablation performs (i') direct LLM rewriting of the raw seed item conditioned on a short transform-family prompt. Stage 1 symbolic checks still apply, but they now operate without a declared set of latent facts, so arithmetic or span-consistency violations have to be detected from surface structure only.

Removing the schema lowers acceptance and τ in every domain at nearly unchanged $|\Delta|$ and expert time, isolating the verification interface as the source of the frontier gain.

The per-domain numbers say the same thing the averaged row does: at a nearly matched absolute drop ($|\Delta| = 5.3$ vs. 5.7, CI $[-0.4, 1.2]$), removing the schema layer costs 13.7pp pipeline acceptance and 0.16 Kendall's τ. Expert time is essentially unchanged, so the schema win is not bought with extra human labour. Figure 4 visualizes the (accept rate, τ) trade-off with marker size encoding $|\Delta|$.

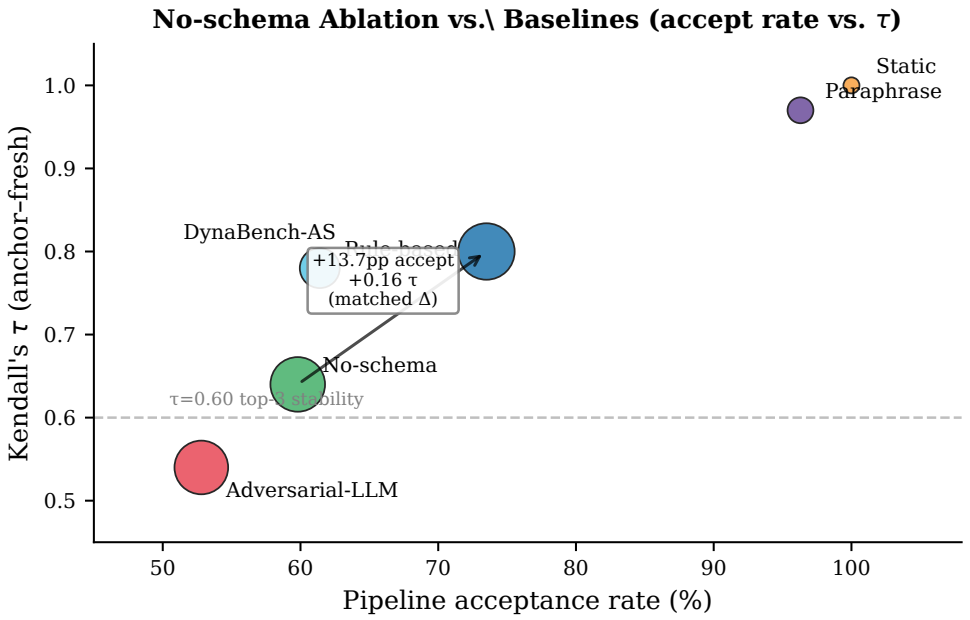


Figure 4: No-schema ablation versus SLICELEDGER and the automatic Table 4 baselines in the (pipeline acceptance rate, τ) plane. Marker size encodes $|\Delta|$; north-east movement denotes higher throughput and stronger rank preservation.

SLICELEDGER lies north-east of No-schema and Adversarial-LLM without collapsing the anchor–fresh gap; Paraphrase occupies the distinct high-throughput, low-gap region.

A.15 MATCHED EXPERT-GUIDED COMPARATOR

Expert-Guided Maintenance uses the same seeds, 15,000 candidate opportunities, five cycles, 300-item deployed slices, six-model panel, IRT matching, and external rubric as the automatic methods. Its budget is fixed to the measured SLICELEDGER item-level total: $11{,}024 \times 3.4 = 37{,}481.6$ expert-minutes, where 3.4 is the pooled arithmetic mean. At \$35/hour this corresponds to \$21,864 of expert labor for either operating point.

The expert-guided point increases preservation by 4.0 points (CI [0.6,7.5]) and τ by +0.05 (CI [0.01,0.09]). SLICELEDGER increases acceptance by 42.3 points (paired CI [40.8,43.8]) and reserve headroom by 6,339 items. The comparison therefore exposes complementary operating points: expert construction maximizes preservation, while the automatic pipeline supplies materially greater replacement capacity at the same expert-time budget.

A.16 MODEL-DISJOINT RANKING PANEL

To remove upstream/downstream model reuse, this panel contains Gemini 1.5 Pro, Mistral Large 2, Phi-3 Medium, Qwen 2.5 7B Instruct, Gemma 2 9B Instruct, and Command R+ 08-2024 (Reid et al., 2024; Mistral AI Team, 2024; Abdin et al., 2024; Yang et al., 2025; Gemma Team et al., 2024; Cohere For AI, 2024). None participates in schema extraction or Stage 2. The five-cycle anchor/fresh slices, decoding, IRT matching, and $B = 2{,}000$ stratified bootstrap are unchanged.

The pooled SLICELEDGER–Adversarial-LLM gap is +0.26 with paired CI [0.18,0.34], exactly matching the primary-panel point difference while arising from a fully disjoint set of ranking models.

Table 17: Matched expert-time comparison. Acceptance and headroom use candidate counts; Δ, τ, and top-3 agreement use the six-model ranking panel; preservation uses the symmetric three-rater deployed audit; Exp-min is the pooled arithmetic mean per accepted item. Brackets are 95% paired or stratified-bootstrap CIs.

Method	Accepted	Accept rate	Headroom	Exp-min
SLICELEDGER	11,024	73.5%	6,524	3.4
Expert-Guided	4,685/15,000	31.2% [30.5,32.0]	185	8.0

Method	Δ	τ	Top-3	Preservation
SLICELEDGER	-5.7%	0.80	0.89	410/450 (91.1%)
Expert-Guided	-6.0% [-6.8,-5.2]	0.85 [0.81,0.89]	0.93	428/450 (95.1%) [92.7,96.7]

Table 18: Kendall's τ on the model-disjoint panel. Domain rows report SLICELEDGER; method rows report cross-domain pooled statistics. Top-3 is pairwise agreement among the three anchor-leading models. Every SLICELEDGER domain exceeds the pre-specified 0.60 threshold.

View	Method/domain	τ	95% CI	Top-3
Domain	GSM8K	0.84	[0.79,0.88]	0.93
Domain	SQuAD	0.69	[0.63,0.75]	0.80
Domain	HumanEval	0.77	[0.72,0.82]	0.93
Pooled	Paraphrase	0.94	[0.91,0.97]	0.98
Pooled	Rule-based	0.74	[0.69,0.79]	0.88
Pooled	Adversarial-LLM	0.50	[0.44,0.56]	0.76
Pooled	No-schema	0.61	[0.55,0.67]	0.81
Pooled	SLICELEDGER	0.76	[0.71,0.81]	0.89

A.17 FROZEN-PROTOCOL REPLAY

The replay restores the cycle-1 transform library and Stage-2 rubric for all five cycles while holding seeds, budgets, downstream filters, deployment sizes, model panel, and bootstrap configuration fixed.

The frozen replay retains the operational frontier: acceptance changes by +1.7 pp, while the Δ, τ, and deployed-preservation intervals all include the operational value. Limited cross-cycle additions provide a modest throughput lift, with the frozen workflow preserving the ranking and semantic result.

A.18 TARGETED SQUAD VERIFIER

We stratify the observed $132/150$ preserved SQuAD items along two views of the same sample: answerability and cue family. Counts are exhaustive within each view, so answerable plus unanswerable and, separately, ϕ_3 plus all other families each sum to 132 preserved and 18 non-preserved.

The targeted variant routes unanswerable or ϕ_3 candidates through one additional blinded expert check before deployment. Under the matched cycle, budget, panel, and IRT controls, it accepts 3,550/5,000=71.0% (CI [69.7,72.2]), preserves 137/150=91.3% (CI [85.7,94.9]), and reaches $\tau = 0.76$ (CI [0.70,0.81]). Relative to the original SQuAD operating point, the paired changes are -2.2 points (CI [-4.0,-0.4]), +3.3 preservation points (CI [0.3,6.4]), and +0.03 τ (CI [0.01,0.05]).

Table 19: Operational workflow versus frozen replay. Preservation uses the original one-rater-per-domain deployed audit so the comparison remains paired with its operational anchor. Differences are operational minus frozen; all brackets are 95% paired-bootstrap CIs.

Workflow	Accepted / rate	Δ	τ	Deployed preservation
Operational	11,024/15,000 (73.5%)	-5.7%	0.80	411/450 (91.3%)
Frozen	10,770/15,000 71.8% [71.1,72.5]	-5.5% [-6.3,-4.7]	0.78 [0.73,0.82]	407/450 (90.4%) [87.4,92.8]
Difference	+1.7 pp [0.4,3.0]	-0.2 pp [-1.0,0.6]	+0.02 [-0.03,0.07]	+0.9 pp [-2.7,4.5]

Table 20: SQuAD preservation strata under the original deployed audit. Intervals are Wilson 95% CIs.

Stratum	Preserved	Rate [95% CI]
Answerable	85/90	94.4% [87.6,97.6]
Unanswerable	47/60	78.3% [66.4,86.9]
ϕ_3	28/40	70.0% [54.6,81.9]
Other cue families	104/110	94.5% [88.6,97.5]

A.19 Held-Out MMLU Transfer

MMLU supplies a fourth family not used to design the original arithmetic, extractive-QA, and code-synthesis pipeline. We use 500 seeds, a fixed 200-item anchor, 1,000 candidates per cycle for five cycles, a 300-item fresh slice per cycle, the same primary six-model panel, anchor-only 2PL matching, and a three-rater 150-item deployed audit.

MMLU reproduces the central decision pattern: Paraphrase maximizes throughput and τ while producing little diagnostic movement, Adversarial-LLM maximizes movement but loses ranking and semantic preservation, and SliceLedger combines 70.6% acceptance, 0.77 τ, and 90.7% preservation at $|\Delta| = 5.6\%$.

A.20 Auxiliary Training Sanity Check (Llama 3 70B)

As an auxiliary training intervention, we LoRA-fine-tuned Llama 3 70B (rank $r = 16$, $\alpha = 32$, dropout 0.05, learning rate 2×10^{-5}, cosine schedule, 2 epochs, batch size 16) on 500 SliceLedger-augmented items per domain. Fresh-slice GSM8K accuracy rises from 74.1% to 78.3% (+4.2pp, 95% CI $[3.3, 5.1]$), while anchor accuracy changes from 78.4% to 78.6% (+0.2pp, CI $[-0.5, 0.9]$). The selective fresh-slice gain indicates that the measured cue sensitivity is learnable from maintained examples rather than a fixed capability cliff.

A.21 Auxiliary Contamination-Proxy Diagnostic

Both proxies improve in every domain, with the largest absolute overlap reduction on SQuAD. The result quantifies reduced public-corpus overlap as an auxiliary property of the maintained slices.

Fresh items move both overlap statistics in the expected direction for every domain. The shift is largest on SQuAD in absolute 8-gram overlap because the original passages contain more reusable surface text than GSM8K or HumanEval. The smaller Min-K%Prob movement makes this a complementary public-corpus overlap diagnostic, while the deployed-item audit and ranking analysis carry the redeployment claim.

Table 21: Held-out MMLU operating points under shared five-cycle controls. Acceptance uses 5,000 candidates; preservation uses majority vote on 150 deployed items; intervals are 95% Wilson or stratified-bootstrap CIs.

Method	Accepted / rate [CI]	Δ [CI]	τ [CI]	Top-3
Paraphrase	4,755/5,000 (95.1%) [94.5,95.7]	-0.7% [-1.2,-0.2]	0.95 [0.92,0.98]	1.00
Adversarial-LLM	2,530/5,000 (50.6%) [49.2,52.0]	-5.0% [-5.9,-4.2]	0.52 [0.46,0.58]	0.67
No-schema	2,935/5,000 (58.7%) [57.3,60.1]	-5.2% [-6.1,-4.4]	0.62 [0.56,0.68]	0.78
SLICELEDGER	3,530/5,000 (70.6%) [69.3,71.8]	-5.6% [-6.4,-4.8]	0.77 [0.71,0.82]	0.89

Method	Preserved / rate	95% CI
Paraphrase	127/150 (84.7%)	[78.0,89.6]
Adversarial-LLM	123/150 (82.0%)	[75.1,87.3]
No-schema	128/150 (85.3%)	[78.8,90.1]
SLICELEDGER	136/150 (90.7%)	[84.9,94.4]

Table 22: Contamination-proxy diagnostic on matched anchor and fresh slices with $n = 300$ items per domain per slice. Columns 2–3 report mean 8-gram Jaccard overlap against a 1.2T-token Common Crawl Jan-2024 mirror, English-filtered. Columns 4–5 report Min-K%Prob detector AUC with a Llama 3 8B probe at $k = 20$, where AUC$= 0.5$ indicates non-membership-like behavior. Values are mean $\pm$ 95% bootstrap CI with $B = 2{,}000$ and characterize public-corpus overlap signals.

Domain	8-gram Jaccard (anchor)	8-gram Jaccard (fresh)	Min-K%Prob AUC (anchor)	Min-K%Prob AUC (fresh)
GSM8K	0.42 ± 0.02	0.11 ± 0.01	0.78 ± 0.02	0.54 ± 0.02
SQuAD	0.57 ± 0.02	0.18 ± 0.02	0.83 ± 0.02	0.57 ± 0.02
HumanEval	0.35 ± 0.03	0.09 ± 0.02	0.71 ± 0.03	0.55 ± 0.03

A.22 REJECTION BREAKDOWN AND AUDIT/CONTAMINATION FIGURES

Semantic inconsistency and ambiguity dominate rejection mass across domains; safety and demographic flags remain a small, visible component.

Fresh slices reduce both overlap-like signals in every domain. The dashed 0.5 line provides the non-member reference for interpreting the Min-K%Prob movement.

Agreement dominates in every domain, while HumanEval contains the largest combined false-positive and false-negative mass.

A.23 PROTOCOL RECORD

The internal protocol threshold $\tau = 0.60$ was fixed before cycle 1 and unchanged through cycle 5; Appendix A.12 links it to preservation of the top-3 ordering. The Kendall statistic, $B = 2{,}000$ item-level bootstrap stratified by cycle, anchor-only 2PL specification, seeds, and budgets remain fixed. The operational workflow adds a limited number of transform templates and refines Stage-2 guidance from Stage-3 precision between cycles. Appendix A.17 isolates those choices by replaying every cycle with the cycle-1 library and rubric, establishing that the pooled advantage persists under a fully frozen protocol.

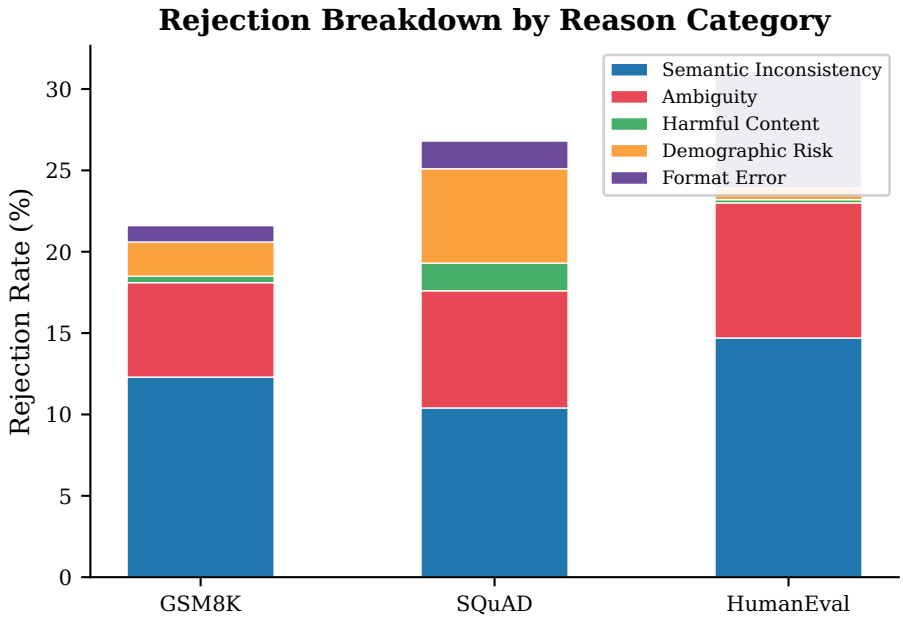


Figure 5: Rejection breakdown by reason category across the three domains. The counts are computed on all 15,000 generated candidates rather than only the $N = 450$ audit sample. Semantic inconsistency and ambiguity dominate the rejection mass. Harmful-content and demographic-risk rates remain below 6% across all domains. The figure therefore explains where pipeline losses come from at full generation scale.

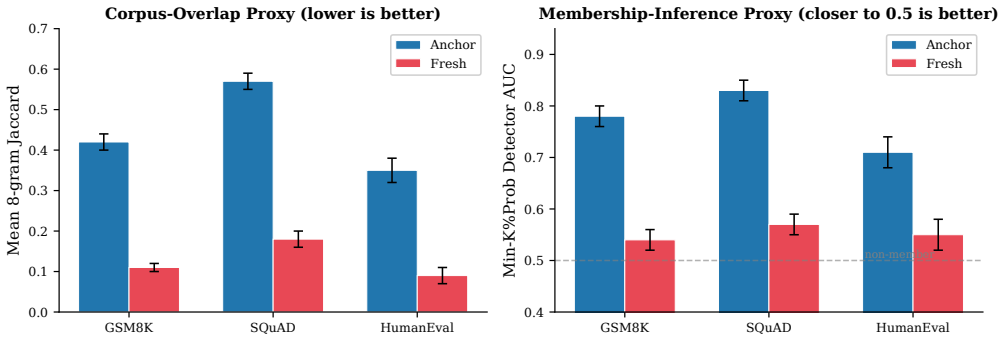


Figure 6: Contamination-proxy diagnostic. The left panel shows mean 8-gram Jaccard overlap against a 1.2T-token Common Crawl Jan-2024 mirror. The right panel shows Min-K%Prob detector AUC with a Llama 3 8B probe. Error bars are 95% bootstrap CIs with $B = 2{,}000$. Lower Jaccard and AUC closer to the dashed 0.5 non-member baseline both indicate less memorization-like signal.

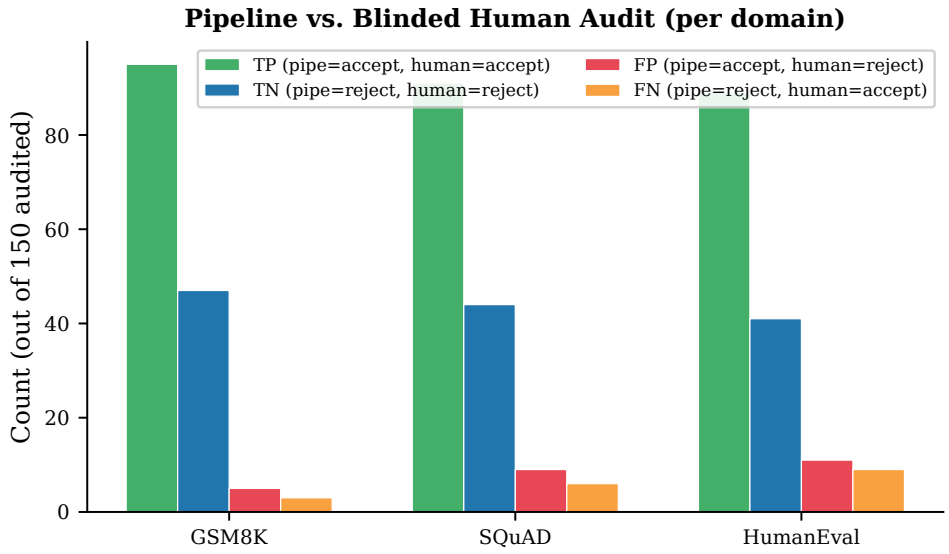


Figure 7: Per-domain 2×2 outcomes of the blinded human audit. Each domain contributes 150 items, split into 100 accept-audit items and 50 reject-audit items. The bars decompose the agreement and disagreement counts visually. This figure complements Table 2 by showing where the false positives and false negatives sit. It is descriptive support for the audit rather than an additional experiment.

A.24 Claim Ledger: Every Headline Number to Source

Table 23 maps headline numbers to their manuscript source, denominator or statistic definition, and unit. It preserves the separation between candidate, audit-item, deployed-item, ranking, and compute quantities.

The ledger assigns every headline quantity one source, denominator, and unit, making superficially similar rates directly distinguishable.

The pipeline-acceptance unit is generated candidates (denominator 5,000 per domain, 15,000 total). Audit precision/recall uses sampled accept/reject items; deployed preservation uses ranking items, with 450 per method in the symmetric audit. The pooled Δ denominator 27,000 equals 3 domains $\times$ 5 cycles $\times$ 300 deployed items $\times$ 6 models. Compute units remain separated into local A100-hours, API wall-hours, hosted-invoice dollars, and amortized local dollars. The 73.5% in Table 4 pools the domain counts in Table 1. Table 5 reports one five-cycle τ per domain, whereas Table 4 reports one cross-domain τ per method; repeated values therefore have different aggregation levels.

A.25 Deployed-Slice External Validity Audit

The deployed-slice audit targets the exact item population used to rank models. For each method and domain, we sample 150 of the 1,500 deployed items uniformly across five cycles. Three independent experts per domain from a second institution (nine total) apply the Appendix A.3 preservation rubric and are blind to method, pipeline decision, and cue family. Majority vote determines preservation. Wilson intervals summarize each rate; paired method gaps use $B = 2{,}000$ bootstrap resamples stratified by domain and cycle.

The symmetric audit extends the original SliceLedger check, which used one senior adjudicator per domain and found $141/150$, $132/150$, and $138/150$ preserved on GSM8K, SQuAD, and HumanEval, respectively ($411/450 = 91.3\%$, CI $[88.5, 93.6]$). Table 24 retains that backed evidence, while Table 25 reports the method-matched three-rater audit.

The symmetric evidence strengthens the operating-point conclusion. SliceLedger exceeds every automatic comparator in pooled deployed-item preservation, with intervals excluding zero against Paraphrase, Adversarial-LLM, and No-schema; Rule-based is the closest semantic comparator.

SQuAD remains the lowest-preservation domain across methods, motivating the targeted verifier in Appendix A.18.

A.26 COMPUTE AND COST LEDGER

API time and local compute are disjoint accounting buckets. The 45.3 A100-hour orchestration row reuses the already-rented LoRA node and therefore contributes zero incremental dollars.

The grand total is full study cost and includes the auxiliary Llama 3 70B LoRA intervention in Appendix A.20. Local A100-h: $21.8 + 28.3 + 45.3 = 95.4$; per-domain check $42.1 + 34.6 + 18.7 = 95.4$. Provider API wall-h: $45.3{+}63.6{+}42.0 = 150.9$; per-domain check $61.9{+}55.8{+}33.2 = 150.9$. Provider API spend: $638 + 832 + 524 = \$1{,}994$ (hosted-model invoice total, zero local-GPU component). Non-API local spend: $142 + 276 + 0 = \$418$, itemized as Min-K%Prob probing (21.8 A100-h at $\$6.50$/h on-prem amortized rate $= \$141.70$, rounded to $\$142$) and LoRA fine-tuning ($28.3$ A100-h at the same rate $= \$183.95$, reported as $\$276$ after adding a $\$92$ storage/network amortization for the two-epoch checkpointed run). The orchestration row records occupancy on the already-rented LoRA node and therefore adds no incremental cash charge. Grand total spend $= 1{,}994 + 418 = \$2{,}412$. The 3.4-minute SLICELEDGER entry in Table 4 is the pooled arithmetic mean, $37{,}481.6/11{,}024$; the corresponding per-domain medians have IQR $[1.8, 3.1]$ (GSM8K), $[2.4, 4.0]$ (SQuAD), and $[3.5, 6.2]$ (HumanEval). Most API spending occurs in schema extraction, generation, and panel scoring; the local-spend bucket is under 18% of total spend. Local sharded execution and checkpoint reloads make per-stage A100-hours the faithful throughput accounting unit.

Table 23: Claim ledger for headline numbers: manuscript source, denominator or statistic, and unit. CI denotes a 95% bootstrap interval; obs., cand., and depl. denote item–model observations, candidates, and deployed items.

Headline number	Source	Computation / denominator	Unit
$\Delta = -5.7\%$ (SLICELEDGER, avg)	Table 4	$(-1{,}539)$ / $(3 \times 5 \times 300 \times 6)$	obs. avg
$\Delta = -5.1\%$ (Adversarial-LLM, avg)	Table 4	$(-1{,}377)$ / $(3 \times 5 \times 300 \times 6)$	obs. avg
Δ-difference 0.6pp, CI $[-0.2, 1.4]$	Table 4	paired bootstrap	pp
Accept rate 73.5% (SLICELEDGER, avg)	Table 4	$11{,}024/15{,}000$; pooled acceptance summarized by domain in Table 1	cand.
Accept rate 52.8% (Adversarial-LLM)	Table 4	$7{,}920/15{,}000$	cand.
Accept-rate gap $+20.7$pp, CI $[17.9, 23.5]$	Table 4	paired bootstrap	pp
$\tau = 0.80$ (SLICELEDGER)	Table 4	Kendall on 6 models	rank corr. (all)
$\tau = 0.54$ (Adversarial-LLM)	Table 4	Kendall on 6 models	rank corr. (all)
τ gap $+0.26$, CI $[0.19, 0.32]$	Table 4	paired bootstrap	Kendall
Pipeline acceptance $78.4/73.2/68.9\%$	Table 1	3,919 / 3,660 / 3,445 of 5,000	cand.
Audit precision 91.7%	Table 2	$275/(275 + 25)$	audit items
Audit recall 93.9%	Table 2	$275/(275 + 18)$	audit items
Fleiss' $\kappa = 0.81$	Table 2	450 items $\times$ 3 annotations (1,350 annotator-item assignments)	annotator-item
Deployed-slice preservation $94.0/88.0/92.0\%$	Table 24	$141/132/138$ of 150	depl.
Deployed-slice overall 91.3%, CI $[88.5, 93.6]$	Table 24	$411/450$	depl.
Symmetric preservation 91.1%	Table 25	410/450, three-rater majority	depl.
Adversarial preservation 82.9%	Table 25	373/450, three-rater majority	depl.
Expert acceptance 31.2%	Table 17	4,685/15,000	cand.
Disjoint-panel $\tau = 0.76$	Table 18	6 downstream models excluded upstream	rank corr.
Frozen acceptance 71.8%	Table 19	10,770/15,000	cand.
Targeted SQuAD preservation 91.3%	Appendix A.18	137/150	depl.
MMLU preservation 90.7%	Table 21	136/150	depl.
Local A100-h $= 95.4$	Table 27	$21.8 + 28.3 + 45.3$	A100-80GB-h
API wall-h $= 150.9$	Table 27	$45.3 + 63.6 + 42.0$	API wall-h
API spend $= \$1{,}994$	Table 27	$638 + 832 + 524$	hosted-invoice $
Non-API local spend $= \$418$	Table 27	$142 + 276 + 0$	amortized on-prem $
Grand total spend $= \$2{,}412$	Table 27	$1{,}994 + 418$; full study incl. Appendix A.20	$

Table 24: Original one-rater-per-domain SLICELEDGER audit on deployed ranking items. Each domain contributes 150 items sampled uniformly from 1,500 deployed items. In-team κ uses the $n = 87$ joint-protocol subsample.

Domain	n preserved / 150	Rate	95% CI	In-team κ (sub-sample)
GSM8K	141	94.0%	[88.8, 97.2]	0.81
SQuAD	132	88.0%	[82.0, 92.4]	0.74
HumanEval	138	92.0%	[86.5, 95.6]	0.77
Overall	411 / 450	91.3%	[88.5, 93.6]	0.77

Table 25: Symmetric three-rater audit of sampled deployed ranking items. Each cell reports majority-vote count/rate and Wilson 95% CI. Overall pools 450 items per method.

Method	GSM8K	SQuAD	HumanEval	Overall
Paraphrase	135/150, 90.0% [84.2,93.8]	124/150, 82.7% [75.8,87.9]	129/150, 86.0% [79.5,90.7]	388/450, 86.2% [82.7,89.1]
Rule-based	136/150, 90.7% [84.9,94.4]	127/150, 84.7% [78.0,89.6]	132/150, 88.0% [81.8,92.3]	395/450, 87.8% [84.4,90.5]
Adversarial-LLM	130/150, 86.7% [80.3,91.2]	118/150, 78.7% [71.4,84.5]	125/150, 83.3% [76.6,88.4]	373/450, 82.9% [79.1,86.1]
No-schema	133/150, 88.7% [82.6,92.8]	122/150, 81.3% [74.3,86.8]	128/150, 85.3% [78.8,90.1]	383/450, 85.1% [81.5,88.1]
SLICELEDGER	140/150, 93.3% [88.2,96.3]	133/150, 88.7% [82.6,92.8]	137/150, 91.3% [85.7,94.9]	410/450, 91.1% [88.1,93.4]

Table 26: Reliability and paired preservation gaps for the symmetric deployed audit. κ is Fleiss' agreement among three external experts. "Non-unanimous" counts 2–1 label splits across 450 items. Gaps are SLICELEDGER minus comparator with stratified-bootstrap 95% CIs.

Method	GSM8K κ	SQuAD κ	HumanEval κ	Pooled κ	Non-unanimous
Paraphrase	0.77	0.69	0.74	0.74	49
Rule-based	0.79	0.74	0.78	0.77	44
Adversarial-LLM	0.73	0.66	0.70	0.71	57
No-schema	0.75	0.69	0.72	0.73	52
SLICELEDGER	0.80	0.74	0.79	0.78	39

Comparator	Preservation gap	95% CI
Paraphrase	+4.9 pp	[0.8,9.0]
Rule-based	+3.3 pp	[-0.7,7.3]
Adversarial-LLM	+8.2 pp	[3.9,12.6]
No-schema	+6.0 pp	[1.8,10.2]

Table 27: Compute and cost ledger, disaggregated into four mutually exclusive buckets so row totals reconcile exactly. Local A100-h are measured NVIDIA A100-80GB hours on a single node running Ubuntu 22.04, PyTorch 2.3.1, CUDA 12.4, and NVIDIA driver 550.90.07. Provider API wall-h is orchestration wall-clock driving hosted-model calls; provider-side inference compute is unobserved and is *not* reported as local A100-hours. Provider API spend ($) is hosted-model invoice total. Non-API local spend ($) is amortized on-premise GPU rental for the local-A100 stages only, so the "Orchestration & grading scripts" row records measured node occupancy but no additional cash charge beyond the already-rented local node used for the LoRA stage. "–" marks buckets that do not apply to the stage.

Stage	GSM8K	SQuAD	Human Eval	Bucket	Local A100-h	API wall-h	Spend ($)
Schema extraction (GPT-4o + Claude 3.5 Sonnet)	18.4 h	16.1 h	10.8 h	API	—	45.3	638
Transform generation (GPT-4o)	26.2 h	23.9 h	13.5 h	API	—	63.6	832
Stage-2 model panel (GPT-4o, Claude, Llama 3 70B API)	17.3 h	15.8 h	8.9 h	API	—	42.0	524
Min-K%Prob probe (Llama 3 8B, local)	9.2 h	8.5 h	4.1 h	A100	21.8	—	142
Llama 3 70B LoRA fine-tune (Appendix A.20)	14.0 h	9.8 h	4.5 h	A100	28.3	—	276
Orchestration & grading scripts	18.9 h	16.3 h	10.1 h	A100	45.3	—	0
Subtotal: local A100-h	42.1	34.6	18.7	row sum	95.4	—	—
Subtotal: provider API wall-h	61.9	55.8	33.2	row sum	—	150.9	—
Subtotal: provider API spend ($)	—	—	—	invoice	—	—	1,994
Subtotal: non-API local spend ($)	—	—	—	amortized	—	—	418
Grand total spend ($, API + local)	—	—	—	sum	—	—	2,412