ANSWER-KEY DRIFT: WHICH BENCHMARK CONCLUSIONS SURVIVE A REVISED KEY

Anonymous authors
Paper under double-blind review

ABSTRACT

When a benchmark is revised, maintainers must decide which published model comparisons still hold. We study this on fixed model outputs, separating changes to answers, scored items, scoring rules and reference actions across twelve archived revision series, a text-to-SQL release and a 2026 tool-use benchmark revision. Headline results can stay put while specific conclusions move. On fixed arithmetic responses from 60 models, one answer-and-item update reverses 10, 9, 10 and 8 of 1,770 comparisons under four scoring rules, and 1,748 comparisons reverse under none of them; yet the four rules yield three distinct reversal sets, with no pair common to all four. On WorkBench, with scorer, tools and state fixed, updated reference actions for 9 of 679 tasks leave the champion and the top-five set unchanged but rewrite 179 of the 216 affected success judgments and reverse 4 of 276 comparisons. The revision's footprint also says what can be reused: a budget certificate that uses only the number of changed tasks certifies 252 of the 276 comparisons from the old scores alone, so all four reversals lie among the other 24. Replacing text-to-SQL references reverses at least six of 120 comparisons under set-equality scoring for every completion of failed evaluations. Across the archived series we express displacement in item-sampling units, locate where edits land, and compare budget rules with same-information baselines.

1 INTRODUCTION

After updating a benchmark, a maintainer must decide which published model comparisons can be reused and which need fresh evidence. Scoring and harness choices are known to reshuffle leaderboards (Alzahrani et al., 2024; Parupudi, 2026). We show that a single revision can leave the headline results in place while changing which specific comparisons hold, and that the revision itself can tell a maintainer which comparisons are safe to reuse. Consider an arithmetic benchmark with stored responses from 60 models. Revising its answers and retained items reverses 10, 9, 10 and 8 of its 1,770 comparisons under four scoring rules, and 1,748 comparisons reverse under none of them. The flagged pairs, however, depend on the rule: 22 pairs reverse under some rule, none under all four, and strict and flexible scoring share one. On the tool-use benchmark WorkBench (Styles & Miller, 2026), with scorer, tools and state fixed, updated reference actions for 9 of 679 tasks leave the champion and the top-five set intact, yet rewrite 179 of the 216 affected success judgments and reverse 4 of 276 comparisons. Knowing only that nine tasks changed, a budget certificate certifies 252 of those comparisons from old scores alone, so every reversal lies among the other 24.

We call this movement *answer-key drift*. The term covers changes to target answers, scored items and judging protocols, while the stored model outputs stay fixed. Our comparison records identify those outputs, the two keys, the scored items, the scoring protocol and original margin, with a permitted revision class when protecting against unknown changes. Sampling uncertainty and key revision are separate perturbations: one changes which items are sampled under a fixed key; the other changes how stored outputs are evaluated. We express the second in units of the first.

The effects depend on the operation and source. The original MMLU/GSM8K correction panels have a median displacement of 0.37 sampling units; SQL reference changes include a reversal whose original margin remains above 1.23 units under all unknown completions. SWE-bench curation moves comparisons by 6.58 and 2.75 units against same-size random subsets. Extraction-induced reranking is already documented (Jo et al., 2025; Kydlicek et al., 2025), as is fixed-output target sen-

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.

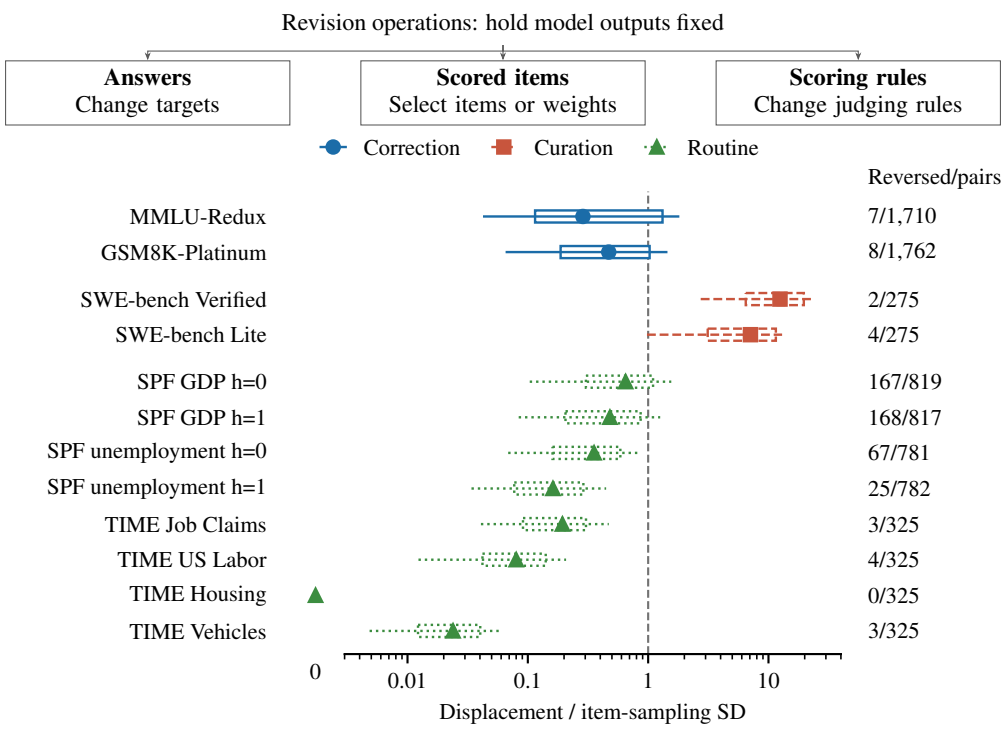


Figure 1: Revision operations on fixed outputs (top) and how far each archived series moved its comparisons, in item-sampling units (bottom, grouped by source rather than operation; median, interquartile box, 10th–90th percentile whiskers). Right: reversed comparisons. Housing has no changed cell and sits at zero; SWE-bench uses the full-task spread, Figure 2 a size-matched one.

sitivity in memory retrieval (Panthi & Abdelfattah, 2026). Our focus is the scale and decomposition of historical changes and the evidence needed to reuse a comparison.

The paper makes three contributions.

1. We measure how far revisions move stored comparisons across twelve archived series and a text-to-SQL release on a common item-sampling scale, keeping every failed evaluation.
2. We separate curation, relabelling, scoring and reference-action changes on fixed outputs, and show that the comparisons a revision overturns depend on the scoring rule even when the number overturned barely changes.
3. We give exact pre-revision budget certificates and completed-revision checks, replay them on the archived series against same-information baselines, and show on WorkBench, under fixed execution, that a certificate built from the number of changed tasks certifies 252 of 276 comparisons from old scores alone.

2 Measuring revisions and checking reuse

Fix two models' outputs, an original key, a scoring rule and normalized nonnegative item weights. Their per-item score or loss difference is g_i and their margin is $d = \sum_i w_i g_i$; either orientation is valid if used consistently at both endpoints. Primes denote the revised quantities. A strict reversal means $dd' < 0$, and a tie at either endpoint is recorded separately. When removal changes the support, the weighting rule is fixed and weights are renormalized over retained items.

For n uniformly weighted original items, we use the empirical item-bootstrap standard deviation and standardized displacement

$$s^2 = \frac{1}{n^2}\sum_{i=1}^{n}(g_i - d)^2, \qquad U = \frac{|d' - d|}{s}. \tag{1}$$

This closed-form spread measures resampling items under the original key. Original ties are excluded from strict-order summaries, every eligible pair in the main collection has positive spread, and zero-spread cases in additional diagnostics keep their declared conventions (Appendix N).

Worked example. On a five-question test, suppose the fixed score differences are $(1, 1, 1, 0, 0)$. The original margin is 0.6 and its item-sampling standard deviation is 0.2191. Allowing one changed answer gives a worst revised margin of 0.2 and displacement 0.4: the sign is retained. Allowing two changes gives a worst margin of -0.2 and displacement 0.8: the rule abstains. These correspond to 1.826 and 3.651 sampling units.

Procedure. Record the fixed outputs, key, support and margin. Specify allowed edits and their budget, then compute the largest adverse displacement. Retain the sign if the absolute margin exceeds that displacement; otherwise report abstention, together with the sampling interval.

Completed-revision check. On fixed support, write $h_i = g_i' - g_i$, $S = \sum_i w_i h_i$ and $E = \sum_i w_i |h_i|$.

Proposition 1 (Target displacement)**.** *For fixed predictions, support and weights,*

$$d' = d + S, \qquad |S| \le E. \tag{2}$$

Thus $E < |d|$ preserves a nonzero sign. For $g_i = |a_i - y_i| - |b_i - y_i|$, let a_i, b_i be the predictions, P_i the projection onto the interval between them, and $\sigma_i = \operatorname{sign}(b_i - a_i)$. Then

$$h_i = 2\sigma_i\{P_i(y_i') - P_i(y_i)\}, \qquad E \le 2\sum_i w_i \min\{|y_i' - y_i|, |a_i - b_i|\}. \tag{3}$$

The projection counts the part of a target change that affects the comparison. A large revision outside the interval between predictions can leave their loss difference unchanged. Appendix B proves the identity and handles removal. This check uses both keys; it quantifies the completed revision.

Pre-revision budget certificate. Declare a revision class $\mathcal{V}$ using the original evaluation. Its exact adverse displacement is

$$D = \max_{v' \in \mathcal{V}} \operatorname{sign}(d)\{d - d(v')\}. \tag{4}$$

For the finite categorical classes and compact boxes used here, a strict sign survives every allowed revision exactly when $D < |d|$. Equality permits a tie.

Proposition 2 (Absolute-loss box budget)**.** *Let $\ell_i(y) = |a_i - y| - |b_i - y|$, $d = \sum_i w_i \ell_i(y_i) \neq 0$, and let every target independently satisfy $|y - y_i| \le \rho_i$. Then*

$$D = \sum_i w_i \max_{|y - y_i| \le \rho_i} \operatorname{sign}(d)\{\ell_i(y_i) - \ell_i(y)\}. \tag{5}$$

Each maximum is attained at a box endpoint. The exact value obeys $D \le 2\sum_i w_i \min\{\rho_i, |a_i - b_i|\}$.

Separate boxes let every item attain its adverse endpoint simultaneously. For categorical accuracy, admissible edits assign one common key to both models or remove an item (Appendix B.1). A leader-favouring disagreement can lose two score units by relabelling; a neutral disagreement can lose one. Sorting the legal adverse gains gives the smallest edit count that destroys the strict sign (Appendix L.2). Invalid predictions remain incorrect under every key. The exact displacement formula and the distinction between a tie and a strict reversal are in Appendices B and L.1.

3 How far a revised key moves a conclusion

In the original twelve-series collection, corrected language benchmarks move comparisons less than curated task sets on the common item-sampling scale (Table 1). The collection contains MMLU-Redux, GSM8K-Platinum, two SWE-bench curations, four Survey of Professional Forecasters series and four TIME economics tasks (Gema et al., 2025; Vendrow et al., 2025; Jimenez et al., 2024). Both keys use the same scorer for each language-benchmark contrast. MMLU comparisons require identical rendered prompts, including ordered choices. Its retained support follows our declared annotation-status policy; the exclusions are analysis choices rather than deletion instructions from the source annotations. The numerical series hold issued forecasts fixed while official targets change (Croushore & Stark, 2001; Qiao et al., 2026). Analysis exposure timing is summarized in Appendix A; survey implementation corrections appear in Appendix C.1. Other analysis changes accompany the affected results.

Across $8{,}521$ nonzero comparisons, $4{,}594$ pass their source-specific check and $3{,}927$ do not; all 458 reversals are among those $3{,}927$. One of the twelve series, the housing task, has no changed cell in the analysed window, and excluding it leaves $8{,}196$ comparisons, the same 458 reversals and $4{,}269$ certified. The pooled median displacement is 0.22 sampling units for routine statistical revisions, 0.37 for corrected benchmarks and 9.0 across the 550 non-tied curation comparisons. The last value pools both SWE-bench subsets using the full-task spread, not the size-matched spread. A median is not the whole distribution, and the tail is what reverses conclusions: the 90th percentile of U is 1.78 and 1.41 on the two corrected releases, and the observed corrections move 28.2% of MMLU and 26.2% of GSM8K pairs by at least one sampling standard deviation, against matched controls of 22.2% and 0.4% (Appendix L). These medians summarize pairwise measurements within the archived revision events. Raw units agree where they are comparable: on the shared scale of item fractions, corrected releases move a typical comparison by 0.0040 and the curated subsets by 0.0781 (Appendix S). The sampling unit divides by each comparison's own spread, which falls with the number of scored items and rises with how often the two models differ. In these MMLU/GSM8K panels, all 15 reversals begin with margins below one sampling standard deviation. Item- and subject-resampling sensitivities remain below the curation scale (Appendix N).

A different boundary in text-to-SQL. On a separate release, we hold $1{,}600$ original SQL outputs fixed across reference-query and database versions. Under set equality, changing only the reference on the old database reverses at least six of 120 model pairs on the full hundred-question support, retaining failed executions as unknown. Contextual-SQL leads MAC-SQL by between 6 and 21 percentage points under the old references; replacing references changes the margin to between -16 and -1 points. These intervals include every completion of failed evaluations. This pair begins beyond 1.23 sampling units under every unknown completion; its displacement is at least 1.43 units. The original-margin lower bound exceeds one unit, not the usual 1.645 or 1.96 Gaussian cutoffs. So a reversal need not begin within one sampling unit. Across the same 120 model pairs, the lower bound is 6 on all 100 questions, 15 on the 56 with unchanged question and evidence text, and 3 on the 37 also having unchanged database bytes. These are different supports, not cumulative counts; the strict subset need not reverse the named pair.

Appendix T separates these input strata and gives conservative full-support displacement intervals rather than dropping failures. The full-support median displacement lies within the envelope $[0, 5.24]$. None of the 516 unknown evaluations is a timeout: they are terminal query errors, row-limit failures or unsupported operations under the fixed contract. Sharing identical cached score targets narrows some displacement intervals but does not change any identified pairwise sign or model-selection result in this sensitivity.

Routine survey and statistical revisions account for 437 of the original collection's 458 reversals. The four benchmark series contribute 21. Appendices S and E.1 give the breakdown, Gaussian reference and quarter-block sensitivities.

A changed pairwise ranking need not change the best model in the evaluated collection. All eight events with a common model roster retain the original champion as a revised-key optimum. The other four series lack this endpoint because forecast availability is pair-specific. An MMLU-Pro check reproduces $24{,}064$ archived item scores. Its fixed Hermes–Qwen ordering survives four official-key contrasts. The largest movement is 0.0171 percentage points (Appendices P and O).

Table 1: Twelve series under revised targets and declared support, excluding original ties. The 4,594 certificates use different information: MMLU/GSM8K bound legal edits on known editable items, SWE-bench bounds deletions within the known removed set, and SPF/TIME use completed target changes. All have zero observed reversals. MMLU exclusions follow our annotation-status policy. SWE-bench medians 12.47 and 7.08 use the original 2,294-task spread; size-matched 500- and 300-task spreads give 6.58 and 2.75 (Section 4), each over 275 non-tied pairs. The final column is the 90th percentile over the same comparisons.

Series	Revision	Models	Pairs	Reversed	Certified	Median U (item SD)	U 90th pct.
MMLU-Redux	Correction	59	1,710	7	1,478	0.29	1.78
GSM8K-Platinum	Correction	60	1,762	8	1,558	0.47	1.41
SWE-bench Verified	Curation	24	275	2	219	12.47	23.67
SWE-bench Lite	Curation	24	275	4	178	7.08	12.98
SPF GDP h=0	Routine	44	819	167	0	0.65	1.56
SPF GDP h=1	Routine	44	817	168	0	0.48	1.26
SPF unemp. h=0	Routine	43	781	67	84	0.36	0.85
SPF unemp. h=1	Routine	43	782	25	165	0.16	0.45
TIME Job Claims	Routine	26	325	3	149	0.19	0.47
TIME US Labor	Routine	26	325	4	217	0.08	0.21
TIME Housing	Routine	26	325	0	325	0.00	0.00
TIME Vehicles	Routine	26	325	3	221	0.02	0.06
Total			8,521	458	4,594		

On fixed arithmetic responses, replacing archived flags with our common parser changes 2,699 cells and reverses 123 comparisons. A historical strict rule reproduces all 79,140 flags, but its runtime remains unidentified. Public strict and flexible rules differ on 3,244 cells and reverse 153 of 1,770 pairs under the original key and support.

With each scorer fixed, updating answers and support reverses 10, 9, 10 and 8 pairs; ties are not counted as reversals. This result has two readings. Reversal status agrees across strict and flexible scoring on 1,753 of the 1,770 pairs, and 1,748 pairs reverse under none of the four rules. The flagged pairs, however, depend on the rule: the four rules yield three distinct reversal sets, with 22 pairs in their union and none common to all four, and strict and flexible scoring share one pair. Appendix R.3 gives fixed-support interactions and pair identities; Appendix U separately examines semantic correctness. The multiple-choice substitution changes no score, a negative control rather than a causal format effect. Main tables keep the common parser under both keys (Appendix M).

Which scoring rules may vary together? For the same arithmetic panel, suppose one of the four studied rules applies to every item. Of 1,754 comparisons that are non-ties under the historical candidate, 1,577 retain their strict sign under every complete rule. Choosing the most adverse rule separately for each item retains 1,512, losing 65 comparisons to this broader relaxation. The original champion passes both tests. The public task configuration reports strict and flexible metrics separately, and the historical candidate and our local parser are additional analysis choices, so the four rules form an explicit sensitivity model. Appendix R.2 gives the policy provenance and concrete witnesses.

For 24 WorkBench archives (Styles & Miller, 2026), changing reference actions between two archived targets on $9/679$ identical tasks leaves the champion and the top-three and top-five sets unchanged, but rewrites 179 of the 216 affected success judgments (82.9%) and strictly reverses $4/276$ comparisons. Both targets are scored with the same current tools, state and scorer. The number of changed tasks alone supports reuse: each changed task can move a pairwise margin by at most two successful tasks, so the budget certificate keeps every comparison whose margin exceeds 18 tasks. It keeps 252 of the 276 comparisons from old scores alone, and all four reversals lie among the 24 it leaves open (Appendix X).

3.1 WHERE A REVISION LANDS

In these aligned panels, corrections concentrate where most models get the old answer wrong, but the models often give different wrong answers. Sharing an error location therefore need not leave their comparison unchanged.

Which items a revision touches decides how far it moves a comparison, and the two operations touch them differently. On fixed support, changing the key on a set R moves the margin by $\frac{1}{N}\sum_{i\in R}(g_i' - g_i)$, and every term with $p_{a,i} = p_{b,i}$ is zero: when both models gave the same answer, any key scores them the same way. Removing a set D moves it by $(|D|\, d - \sum_{i\in D} g_i)/(N - |D|)$, so an item answered alike contributes $d/(N - |D|)$ and an item they split up to $(1 \mp d)/(N - |D|)$. Both statements are exact, and we verified them in integer arithmetic on every one of the $5{,}192$ model pairs of the three aligned panels, reproducing each archived margin with no discrepancy. Two of the panels are one MMLU-Redux release under two prompt-identity rules, sharing the 59 models and overlapping in items: one source counted twice.

The identities depend on both the edited locations and the new labels. After seeing the displacement results, we registered a descriptive test of placement that uses original-key information only: for item i let c_i be the fraction of models scored correct under the original key and $v_i = c_i(1 - c_i)$, the quantity for which $2c_i(1 - c_i)M/(M - 1)$ is the probability that two uniformly drawn models differ on that item. The observed edit sets are compared with $10{,}000$ uniform subsets of the same size drawn from the panel (Appendix R).

Figure 3 in Appendix R shows these item distributions.

Where corrections are made. On the three panels, 64.6%, 65.0% and 80.0% of corrected items were answered correctly by at most a fifth of models, against uniform-location medians of 13.1%, 13.0% and 20.0%. The companion original-key disagreement test points the same way; both registered placement tests reject on all three panels at Holm-adjusted $p \leq 0.0007$ (Appendix R).

Failing together is not answering alike. On those same corrected items two uniformly drawn models give the same answer with probability 0.538, 0.541 and 0.175, against 0.561, 0.562 and 0.254 over all items of the panel. Counting exactly, one correction changes the margin of 42.7 and 42.4 per cent of the model pairs on the multiple-choice panels. Where corrections are made is therefore a fact about editors, not an explanation of how far a correction moves a conclusion. The unstratified matched control of Section 4 gives a median displacement of 0.220 against 0.443 at random locations. Conditioning also on subject, legal label transitions and original model performance narrows that contrast: a prespecified coarse sensitivity gives a reference median of 0.238, with a 5th–95th percentile range $[0.194, 0.290]$ containing the observed value. Exact matching fixes 97 of the 99 edit locations. These controls show selection sensitivity and limited scope for relocation (Appendix R.1). Removal placement depends on the source and exclusion policy; the registered test reports one direction on the multiple-choice panels and the other on the arithmetic one (Appendix R).

4 CURATION AND JUDGING ARE DIFFERENT OPERATIONS

Curation moves comparisons beyond the variation expected from retaining fewer items at random. For a fixed model pair, let D_{diff} count disagreeing scores on the full set of N items, and let d be their mean difference. A uniformly selected subset of m items has

$$s_m^2 = \frac{N - m}{m(N - 1)} \left(\frac{D_{\text{diff}}}{N} - d^2 \right). \tag{6}$$

This reference includes the change in sample size. We reconstruct the disagreeing items from archived resolved-task lists, keeping failed tasks as zeros. Relative to this reference, Verified and Lite move the median comparison by 6.58 and 2.75 standard deviations (Figure 2). Crossed model-and-item resampling gives 2.5–97.5 percentile ranges $[4.66, 8.50]$ and $[1.85, 3.90]$.

The discrete reference also supplies an exact conditional test for each comparison. After Holm correction across both subsets and all pairs, $198/276$ Verified and $92/276$ Lite comparisons reject uniform selection. Tests include the original tied pair in each subset; strict-order medians exclude it. The tests establish departure from uniform subset selection (Appendix Q).

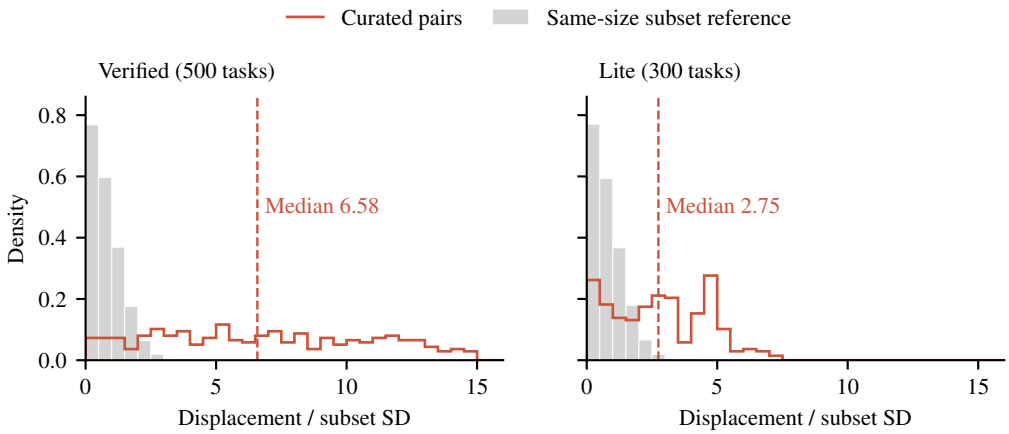


Figure 2: Curated-pair displacement and a size-matched random-subset reference for SWE-bench. Grey densities average the exact pair-specific subset laws over the same nonzero-margin pairs; outlined densities and dashed medians describe the observed curation. The reference is a mixture of pairwise laws, not the sampling distribution of their median.

Separating label changes from item removal. Matched controls hold the exclusion and relabelling counts fixed and apply one shared key to all models (Table 2). On the primary and sensitivity MMLU panels and GSM8K, relabelling-only displacement lies below the random-control median while deletion-only displacement exceeds its 95th percentile: the two operations act in opposite directions against the same reference. Joint-edit controls are mixed.

Table 2: Matched random controls at declared counts. D counts exclusions and C relabelled items; Flips and U_{50} compare observed reversals and median displacement with uniform-location medians and 5–95 percentile ranges. MMLU exclusions follow our annotation-status policy. Joint edits are in Appendix L.1.

Panel	Action	D	C	Flips	Random	U_{50}	Random
MMLU exact	delete	254	0	7	3.00 [1.00, 6.00]	0.253	0.148 [0.115, 0.198]
MMLU exact	relabel	0	99	6	5.00 [2.00, 8.00]	0.220	0.443 [0.362, 0.542]
MMLU WS	delete	263	0	4	3.00 [1.00, 6.00]	0.260	0.149 [0.112, 0.201]
MMLU WS	relabel	0	100	8	5.00 [1.95, 8.05]	0.215	0.442 [0.324, 0.535]
GSM8K	delete	110	0	7	7.00 [4.00, 11.00]	0.458	0.201 [0.164, 0.260]
GSM8K	relabel	0	10	1	1.00 [0.00, 3.00]	0.068	0.159 [0.119, 0.184]

Changing the judging protocol. AlpacaEval regrades fixed outputs (Li et al., 2023). With judge and reference fixed, joint prompt, format, decoding and parser changes move the median 2.91 sampling units, reverse $7/66$ pairs and leave 13 passing the completed-revision check. Changing judge and reference moves 5.29 units and reverses $115/1{,}222$ pairs. For Arena-Hard hard-prompts, raw-score winners change but the official style-controlled winner does not (Appendices J and K).

5 When an old comparison can be reused

Pre-revision guarantee and its coverage condition. A nonzero old sign survives every revision in the declared feasible class when its largest adverse displacement is smaller than the old margin (Equation 4). This is a guarantee for an actual update only when that class contains the update. With valid categorical predictions, uniform weights and unrestricted unit-cost relabelling, strict preservation reduces exactly to $N|d| > 2K$ for edit budget K: in this class the certificate adds no information beyond the margin (Appendix L.2). Numerical rows below use the relaxed box bound $2\sum_i w_i \min\{\rho_i, |a_i - b_i|\}$; exact boxes give a separate comparison.

Retrospective replay of budget rules. Combining a sampling interval with the published touched-count budget on categorical series and historical envelopes on numerical series retains $2{,}506$ of $8{,}521$ comparisons with no observed reversal (Table 16). The interval alone retains $5{,}712$, of which 22 reverse. The budget rule alone retains $2{,}874$, with seven reversals. The two sets of missed reversals are disjoint in this collection. The conjunction excludes GDP and both SWE-bench curations. Table 1 instead uses each realised revision or a known editable set.

What the budget rule contributes at that retention is measurable, and we registered the comparison before running it. Keeping $2{,}506$ different pairs by top $|d|/s$, pooled or by series, also leaves no observed reversal, and on the eight series where the budget rule retains anything the interval alone keeps $4{,}757$ and reverses none (Table 15). All 22 reversals above the conjunction's smallest retained $|d|/s$ of 1.646 lie in series the budget rule declines whole, so its contribution here is that series-level abstention rather than a finer separation inside the series it keeps. The retained identities differ.

The sampling rule requires $|d|/s > 1.6448536$. Its two-sided sensitivity and the hindsight frontiers are in Appendix L; the budget inputs and abstention denominators are in Appendices F and D. Appendix G indexes the numerical records.

Other archived coverage. A budget calibrated on earlier vintages retains more than the historical envelope. On the claims task a split-conformal displacement quantile calibrated on the vintages before the median release date (Vovk et al., 2005; Angelopoulos & Bates, 2021) retains 313 of 325 comparisons at $\alpha = 0.10$, none of which reverses on the later vintages, against 93 for the relaxed historical envelope; three of the 325 reverse in total. A standardised margin computed on the same calibration window selects exactly those 313 comparisons, and discarding comparisons whose sign already changed during calibration retains 322, both with no reversal (Appendix I).

The exact box rule increases coverage at an unchanged budget: at the historical 90th-percentile budget it retains 275 claims comparisons against 93 for the relaxed bound, and for the two unemployment horizons 110 versus one and 312 versus six, with no added pair reversing in this replay. The proof supplies conditional validity; these counts measure the cost of a loose bound on archived inputs (Appendix L). The key's revision horizon also changes the survey counts (Appendix E).

Registered future tests. The first registered checkpoint, as of 2026-09-14, resolves four tasks and $1{,}300$ pair decisions. It produces one reversal, on a pair for which the certificate abstained. Three tasks have no revised cell. The next checkpoint, as of 2026-11-02, and the registered 2026-11-16 and 2027-02-16 resolutions are the prospective tests (Appendix H).

6 Related work

SQL corrections alter questions, references, databases and rankings (Jin et al., 2026a;b); our fixed-output factorial separates reference and database effects while retaining failures. Liu et al. (2026) cross execution oracles for conformal SQL abstention, with expert and same-SQL checks. Their fixed candidate pools may yield different SQL selections; our factorial fixes every submitted query.

Scoring and harness choices change leaderboards (Alzahrani et al., 2024; Parupudi, 2026), and answer extraction and scorer replacement rerank fixed responses (Jo et al., 2025; Kydlicek et al., 2025); we ask which comparisons one revision overturns under each rule. Panthi & Abdelfattah (2026) show fixed-trace target sensitivity with shared-query and semantic controls. Test-set label errors (Northcutt et al., 2021) and model-in-the-loop construction (Phang et al., 2022) can change rankings.

Small-subset ranking tests and task selection (Yauney et al., 2026; Gusev & Zaytsev, 2026), sequential stopping (Arviv et al., 2026), exact unknown-label winner certificates (Kuboyama, 2026), and cost-aware active rating (Angelopoulos et al., 2025) precede this work. Incremental retrieval pools reuse earlier judgments under fixed-label and fixed-list conditions (Nelson et al., 2026).

Real-time data revisions already affect predictive-ability tests (Clark & McCracken, 2009); our object is the finite change in stored comparisons. Test equating and fixed-parameter IRT link score scales across forms and growing suites (Kolen & Brennan, 2014; Habba et al., 2026), rather than bound revisions to stored answers. Human-calibrated noisy judges support statistical tests (Feng et al., 2026); we bound finite revision sets instead. WorkBench maintainers report fixed-model

score changes across versions (Styles & Miller, 2026); we add pairwise, champion and top-k audits under one fixed execution contract. Appendix W details these distinctions and the measurement background.

7 LIMITATIONS

Our panels mostly cover 2023–2024 open models, with 2026 runs from the tool-use archives. Pairs share models and items, so pooled summaries describe this collection, not a population of revisions or model-selection losses, and the pooled reversal rate is not a benchmark-specific estimate. Several analyses followed exposure to the revisions: the scorer census concerns one archive and operations of different sizes; the placement results cover one multiple-choice release under two identity rules and one arithmetic release, and are neither independent replications nor causal effects of curation or of source-family differences; their uniform-location reference does not reproduce Platinum's failure-based screening (Vendrow et al., 2025), the arithmetic correction set has ten items, and MMLU exclusions follow our annotation-status policy rather than source deletion instructions; curation resampling ranges measure sensitivity within the archived models and task memberships, not curator effects.

Budget guarantees hold only when the declared revision class covers the update, and zero reversals among certified comparisons (Table 1, WorkBench) follow from their declared change sets, not from independent risk validation. The archived replay is retrospective: published categorical edit counts are known only after a revision, all four categorical revisions exceed the alternative one-twentieth-item budget, and containment of the numerical envelopes is unverified, so the $1{,}075$ numerically retained comparisons carry no coverage guarantee. A graded reading of the bound does not survive stratification: the budget-to-margin ordering lies between series, not within them (Appendix L.3). A hindsight threshold scan is an outcome-informed reference, not a rule for an unseen revision. Historical quietness can fail: the TIME envelopes retain $1{,}068$ comparisons, of which seven later reverse (seven of the ten TIME reversals), and raising the quantile to the largest historical revision leaves the same failures. Same-information margin rules match the budget conjunction and the conformal claims-task selection, so those retentions carry no method advantage; the conformal calibration vintages are not exchangeable, so no coverage guarantee holds. The resolved checkpoint is descriptive, because its changed task resolves through a vintage older than registration; the registered future tests remain open.

On WorkBench, historical tool versions are unverified and model names are upstream archive labels; the certificate assumes that unchanged tasks keep their scores under the fixed current scorer, tools and state, and its threshold of 18 does not certify the champion, whose original lead is 7. The code-checker replay (Appendix V) leaves many positions unknown; no retention count here, WorkBench's included, establishes cheaper verification or prospective selection, and the four-rule scorer comparison is a conditional static analysis, not a new algorithm. SQL reference replacement includes changed questions, consumed prompt traces are unavailable, one large SQL reversal does not establish a larger typical effect, and extra compute cannot supply contractually undefined scores. AlpacaEval's pair-specific instruction intersections preclude isolated judge or human-preference claims. The MMLU-Pro outputs postdate the early corrections. Semantic truth and historical scorer dependencies remain unverified, AI judgments are not human gold, no independent free-text source confirms the scorer contrast, and traffic archives (Appendix C) require authorised access.

CONCLUSION

Benchmark revisions should be audited at the level of specific conclusions. Record score, pairwise-order and champion changes separately, reuse the comparisons whose margins exceed what the declared change can move, and re-check the rest.

Reproducibility Statement. The supplementary analyses reconstruct the twelve-series tables from per-comparison records, including the aligned MMLU and GSM8K scores. These records supply margins, revision bounds and item moments; the derived-table program recomputes strict decisions, exposure bins and sampling-unit summaries. Reconstructing these statistics from raw outputs is a separate operation.

The GSM8K source reconstruction applies one scorer to the original and revised keys. It has been independently rebuilt in a clean environment from public model records and annotations. The MMLU source reconstruction covers both identity rules, target parsing, scores and finite-key bounds. Its offline rebuild agrees with the recorded results; the separate fresh-download attempt was incomplete because of upstream rate limiting. The other raw analyses require their source-specific archived releases. The traffic case also requires authorized PeMS access. The distributed package contains derived records and analysis programs; model weights and private traffic records are excluded.

The command `python -B run.py all --out NEW_DIR` reconstructs the main comparison table, common noise reference, intersection of the two rules and first checkpoint summary. Each operation also has a separate entry in the package instructions. Figure data are derived from frozen comparison records; the figures use editable PGF. The self-contained paper source builds with `python3 build_paper.py`. The numerical program was run in an empty directory from the archive alone: it finished with no network request and no model run, and reproduced $8{,}521$ comparisons with 458 reversals, $5{,}712$ and 22 for the sampling rule, $2{,}874$ and 7 for the budget rule, $2{,}506$ and 0 for their conjunction, and $1{,}300$ checkpoint decisions with one reversal; its twelve-series output matched all sixty source cells of five count columns, models, pairs, reversed, certified and reversed within the certified set, on every one of the twelve series. Table 1 displays the first four and states the all-zero fifth in its caption. The paper source was extracted into a fresh directory and rebuilt there to the same page count. The additional `revision_controls.py` entry reconstructs containment states, matched-retention and same-calibration-information baselines. The `scorer/scorer_contracts.py` entry tests explicit parsing contracts and rebuilds factorial contrasts from derived cell scores; it does not distribute original generated text or supply human semantic annotations. The separate `shared_structure_supplement.zip` provides `python -B run.py --out NEW_DIR` to check the scoring-policy witnesses, classify archived SQL failures and reconstruct the identity-constrained bounds. It uses derived scores and hashed execution identities; it does not retry failed queries or distribute databases and generated SQL. The paper source includes its tables and figures. Appendix G identifies the supporting numerical records, Appendix A gives exposure timing, and Appendix B contains the proofs. The checkpoint computation in Appendix H has an independent offline reconstruction from archived responses and frozen predictions. Future scheduled resolutions require new data at their registered dates.

Ethics Statement. We analyse public archives, published benchmarks and public survey releases, without collecting new human-subject data or private records. Our criticism concerns published records and released artifacts, not the competence or intent of their authors.

DISCLOSURE OF AI USAGE

The authors directed this research, made all scientific decisions, and take full responsibility for the content of this paper. Under the authors' direction, we used large language model assistants as tools for proposing and refining hypotheses and theoretical framing, formulating mathematical claims and drafting and checking proofs, designing and critiquing methods and experiments, implementing methods, generating synthetic test data, cleaning and reformatting data, translation, qualitative review, interpreting results, searching the literature, and drafting and editing text. The supplementary semantic audit uses AI judgments as its data, as described in its appendix. No AI system is an author.

REFERENCES

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, 2024. doi: 10.18653/v1/2024.acl-long.744.

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

Anastasios N. Angelopoulos, Jacob Eisenstein, Jonathan Berant, Alekh Agarwal, and Adam Fisch. Cost-optimal active AI model evaluation, 2025. URL https://arxiv.org/abs/2506.07949v1.

Ofir Arviv, Kristjan Greenewald, Yotam Perlitz, Hadar Mulian, Michal Shmueli-Scheuer, and Leshem Choshen. Stop guessing when to stop testing: Efficient model evaluation with just enough data. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 871–881. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.43. URL https://aclanthology.org/2026.findings-acl.43/.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models, 2024. URL https://arxiv.org/abs/2405.14782v1.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*, 2020.

Todd E. Clark and Michael W. McCracken. Tests of equal predictive ability with real-time data. *Journal of Business & Economic Statistics*, 27(4):441–454, 2009. doi: 10.1198/jbes.2009.07204. URL https://doi.org/10.1198/jbes.2009.07204.

Dean Croushore and Tom Stark. A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130, 2001.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263, 1995.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hEhxreaLdU.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

Rostislav Gusev and Alexey Zaytsev. Benchmarking on tasks that matter: Dataset selection for preserving model rankings. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817569. URL https://arxiv.org/abs/2606.27997v1.

Eliya Habba, Itay Itzhak, Asaf Yehudai, Yotam Perlitz, Elron Bandel, Michal Shmueli-Scheuer, Leshem Choshen, and Gabriel Stanovsky. Growing pains: Extensible and efficient LLM benchmarking via fixed parameter calibration, 2026. URL https://arxiv.org/abs/2604.12843.

Jenny Y. Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. Dropping just a handful of preferences can change top large language model rankings. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.11847.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

Tengjun Jin, Yoojin Choi, Yuxuan Zhu, and Daniel Kang. Text-to-sql benchmarks are broken: An in-depth analysis of annotation errors. In *Conference on Innovative Data Systems Research*, 2026a. URL https://www.vldb.org/cidrdb/papers/2026/p5-jin.pdf.

Tengjun Jin, Yoojin Choi, Yuxuan Zhu, and Daniel Kang. Pervasive annotation errors break text-to-sql benchmarks and leaderboards. *Proceedings of the VLDB Endowment*, 19(5):931–944, 2026b. doi: 10.14778/3796195.3796206. URL https://www.vldb.org/pvldb/vol19/p931-jin.pdf.

Hwiyeol Jo, Joosung Lee, Jaehone Lee, Sang-Woo Lee, Joonsuk Park, and Kang Min Yoo. Finding answers in thought matters: Revisiting evaluation on large language models with reasoning, 2025. URL https://arxiv.org/abs/2510.14773v1. Version 1, 16 October 2025.

Michael J. Kolen and Robert L. Brennan. *Test Equating, Scaling, and Linking: Methods and Practices*. Springer, 3 edition, 2014. doi: 10.1007/978-1-4939-0317-7. URL https://doi.org/10.1007/978-1-4939-0317-7.

Stefan Krsteski, Charlotte Meyer, Guillaume Allegre, Tony O'Halloran, and Alexandre Sallinen. Messier: A high-resolution corpus for cross-benchmark agent evaluation, 2026. URL https://arxiv.org/abs/2607.25891v2. Version 2, revised 30 August 2026.

Yizhou Kuang. Revision risk in real-time macroeconomic forecasting. *arXiv preprint arXiv:2607.05882*, 2026. doi: 10.48550/arXiv.2607.05882. URL https://arxiv.org/abs/2607.05882v2. Version 2, revised 26 July 2026.

Tetsuji Kuboyama. How many labels does model choice need? certificates and budgets for selective prediction, 2026. URL https://arxiv.org/abs/2609.18622v1.

Hynek Kydlicek, Alina Lozovskaya, Nathan Habib, and Clémentine Fourrier. Fixing open LLM leaderboard with Math-Verify. Hugging Face Blog, February 2025. URL https://huggingface.co/blog/math_verify_leaderboard. Published 14 February 2025; accessed 22 September 2026.

Stefan Larson, Attila Nagy, Sam Desai, Cyrus Desai, Nicole C. Lima, Yixin Yuan, Siddharth Betala, Kaushal K. Prajapati, Jamiu T. Suleiman, Sharad Duwal, and Kevin Leach. Revising rvl-cdip: Quantifying errors and test-train overlap, 2026. URL https://arxiv.org/abs/2606.31446v1. Version 1.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024. URL https://arxiv.org/abs/2406.11939v2.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.

Yitao Li. Who drifted: the system or the judge? anytime-valid attribution in LLM evaluation pipelines. *arXiv preprint arXiv:2606.15474*, 2026. URL https://arxiv.org/abs/2606.15474v1.

Jiamiao Liu, Dewen Qiao, Yu Zhang, and Xuetao Chen. Certified against which oracle? execution labels set the reported risk of conformal abstention for text-to-SQL, 2026. URL https://arxiv.org/abs/2609.25938v1.

LiveCodeBench developers. Handle exit in the autograder. Git commit 15808c76fbd459a73ce6b60d111e56860ab45f6d, 2025. URL https://github.com/LiveCodeBench/LiveCodeBench/commit/15808c76fbd459a73ce6b60d111e56860ab45f6d.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229v1.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640v1.

Ajinkya Mohgaonkar, Lukas Gosch, Mahalakshmi Sabanayagam, Debarghya Ghoshdastidar, and Stephan Günnemann. Exact certification of neural networks and partition aggregation ensembles against label poisoning, 2026. URL https://arxiv.org/abs/2604.11416v1. Version 1; Workshop on Principled Design for Trustworthy AI at ICLR 2026.

Max Nelson, Hanoz Bhathena, Aviral Joshi, and Saket Sharma. Incremental pooled LLM evaluation for cost-effective retrieval model selection, 2026. URL https://arxiv.org/abs/2609.02745v1.

Yannic Neuhaus and Matthias Hein. RePOPE: Impact of annotation errors on the POPE benchmark, 2025. URL https://arxiv.org/abs/2504.15707v1. Extended abstract, Third Workshop on What is Next in Multimodal Foundation Models (MMFM), held with CVPR 2025.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2103.14749. Datasets and Benchmarks Track.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024. URL https://arxiv.org/abs/2404.13076.

Sugam Panthi and Rabab Abdelfattah. Same ranking, different winner: How scoring targets shape LLM memory benchmarks, 2026. URL https://arxiv.org/abs/2605.24060v2. Version 2, revised 26 August 2026.

V. S. Raghu Parupudi. There is no neutral harness: Modern LLM leaderboards are manufactured by config-fragile items, 2026. arXiv:2608.21382.

Sarina Penquitt, Jonathan Klees, Antonia van Betteray, Parssa Jashnieh, Peter Stehr, Matthias Rottmann, and Lars Schmarje. Object detection benchmarks are incomplete: The role of label errors and annotation uncertainty, 2026. URL https://arxiv.org/abs/2609.21822v1. Version 1, 18 September 2026.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Benchmark agreement testing done right: A guide for llm benchmark evaluation, 2024. URL https://arxiv.org/abs/2407.13696v1. Version 1; version 2 is titled Do These LLM Benchmarks Agree? Fixing Benchmark Evaluation with BenchBench.

Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. Adversarially constructed evaluation sets are more challenging, but may not be fair. In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, pp. 62. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.dadc-1.8. URL https://aclanthology.org/2022.dadc-1.8/.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. *arXiv preprint arXiv:2602.12147*, 2026.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3NjnRo6apU.

Olly Styles and Sam Miller. WorkBench revisited: Workplace agents two years on, 2026. arXiv:2606.13715.

Olly Styles, Sam Miller, Patricio Cerda-Mardini, Tanaya Guha, Victor Sanchez, and Bertie Vidgen. WorkBench: a benchmark dataset for agents in a realistic workplace setting, 2024. URL https://arxiv.org/abs/2405.00823v2.

Veerendra Kumar Sunkavalli. LLM judges as raters: A pre-registered audit of severity, halo, reliability, and version instability in LLM essay scoring on public corpora. *arXiv preprint arXiv:2608.29517*, 2026. URL https://arxiv.org/abs/2608.29517v1.

Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel-Lamparth, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Benjamin Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in ai benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2118. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5b44b99d012922011f0be1031a98c4f9-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*, 2025.

Illia Volkov, Nikita Kisel, Tetiana Mishkina, Klara Janouskova, and Jiri Matas. Doomed to reannotate, forever: The imagenet story, 2026. URL https://arxiv.org/abs/2608.13783v1. Version 1.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023. URL https://arxiv.org/abs/2305.17926.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.01574.

Gregory Yauney, Shahzaib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 27389–27412,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2e2960f2fe9e981f33f51c78656e3ca2-Paper-Conference.pdf.

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 *Advances in Neural Information Processing Systems*, 2023.

A STUDY DESIGN AND EXPOSURE TIMING

The archived comparisons and the analyses of those comparisons have different exposure times. Table 3 separates their registrations. All reported MMLU and GSM8K results use the common-scorer MMLU and GSM8K populations described in Appendices N and M.

Table 3: Exposure timing of the reported analyses. An analysis can be registered after an earlier endpoint was observed. Future-checkpoint status depends on the resolved vintage.

Analysis	Information available when its rules were fixed
Archived key comparisons	Models, item admission and pairwise rules were registered before the corresponding prediction or target files were read.
Common-scorer alignment	Historical endpoints were known; identities, parsers and exclusion policies were fixed before the aligned comparisons were computed.
Sampling-unit summaries	Historical margins were known. The common analytical item variance was fixed before the harmonized summaries.
Matched controls	Historical endpoints and edit counts were known; legal shared-key actions, the random stream and every contrast were fixed before the draws.
Historical budget comparisons	Outcomes were already available; the box calculation and sampling/budget intersection are retrospective analyses.
Scheduled checkpoints	Forecasts, cells, weights and historical envelopes were frozen before retrieval. The first checkpoint includes revisions published before registration.
Code-checker secondary replay	After a separate post-exposure one-question diagnostic, all 713 questions and each model's first completion were specified before secondary scores. An input-digest correction preceded both scorer launches.

B PROOFS OF THE REVISION CERTIFICATES

B.1 CATEGORICAL SCORES AND EDITABLE SUPPORT

For zero-one scores, let $d = N^{-1}\sum_i\{s_A(i)-s_B(i)\} \neq 0$ and $\sigma = \operatorname{sign}(d)$. In the editable set R, let P_R contain items favouring the leader, with contribution $\sigma\{s_A(i)-s_B(i)\} = 1$. When relabelling is admissible, Z_R contains neutral items whose fixed answers differ and whose replacement key may favour the challenger; otherwise $Z_R = \varnothing$. The remaining neutral items form Z_R^0. The categorical result assumes that each edited item can take the challenger's answer, independently across items.

Proposition 3 (Exact zero-one displacement)**.** *Under relabelling (admissible when $\mathbf{1}_{\mathrm{rel}} = 1$) and deletion, the largest displacement of the margin toward a reversal over all admissible revisions of R is*

$$E(R) = \max_{\substack{0\le k\le |P_R|,\ j\in\{0,|Z_R^0|\}\\ N-k-j>0}} \Big[\,|d| - \frac{|d|\,N - k - 2\,(|P_R|-k)\,\mathbf{1}_{\mathrm{rel}} - |Z_R|}{N-k-j}\Big], \tag{7}$$

and the sign of d is preserved under every admissible revision of R if and only if $E(R) < |d|$.

Relabelling a leader-favouring item changes its contribution from $+1$ to -1; deleting it removes one from both the numerator and the count. The proof and exact-arithmetic checks are in Appendix B. Missing GSM8K answer parses remain incorrect under every key, so that panel uses direct enumeration of its finite shared-key domain (Appendix M). For survey targets, Proposition 1 supplies the absolute-loss bound.

In the proofs below, the index c denotes an item, as does i in the main text; $d_v = d$ and $d_{v'} = d'$ denote the margins under the two keys.

Proof of Proposition 1. Subtracting the two weighted margins gives $d_{v'} - d_v = S$; the triangle inequality gives $|S| \le E$. The strict reversal and tie statements follow by locating $d_v + S$ relative to zero. The inequality $|S| \le E < |d_v|$ proves the exposure certificate.

For $a_c \le b_c$, the absolute-loss difference is constant outside $[a_c, b_c]$ and linear with slope two inside it. With $m_c = (a_c + b_c)/2$, it equals

$$|a_c - y| - |b_c - y| = 2\sigma_c\{P_c(y) - m_c\}. \tag{8}$$

Reversing the predictions multiplies this expression by $\sigma_c = -1$, and equal predictions make both sides zero. Projection onto an interval is nonexpansive and has range diameter $|a_c - b_c|$, proving the MAE identity and bound in equation 3. □

Proof of Proposition 2. For $a_c < b_c$, $\ell_c(y)$ equals $-(b_c - a_c)$ for $y \le a_c$, $2y - a_c - b_c$ on $[a_c, b_c]$ and $b_c - a_c$ for $y \ge b_c$; for $a_c > b_c$ the signs reverse. So ℓ_c is monotone and piecewise linear with slope at most 2 in absolute value, its extremes over $[y_c - \rho_c, y_c + \rho_c]$ sit at the endpoints, and the largest adverse move in cell c is the maximum in equation 5. The cells are constrained separately, so choosing the adverse endpoint in every cell at once is admissible and moves d by exactly D, and no admissible revision moves it further. If $D < |d|$ the moved margin keeps its sign; if $D > |d|$ the simultaneous endpoint choice crosses zero, and by continuity of ℓ_c a revision with the opposite sign exists; $D = |d|$ gives a tie. The relaxation follows from the slope bound: the displacement from the centre of a box of half-width ρ_c to either endpoint is at most $2\rho_c$, and it is also at most the total range $2|a_c - b_c|$. The range across the entire box can be as large as $4\rho_c$; it is the centre-to-endpoint displacement that enters the relaxation. □

For the exposure factorisation, let $\mathcal{R}$ contain the items whose targets change, and let $W_R = \sum_{c\in\mathcal{R}} w_c$. Set $M_R = E/W_R$ when $W_R > 0$ and zero otherwise. Set $\alpha = S/E$ when $E > 0$ and zero otherwise. Then

$$E = W_R M_R, \qquad S = W_R M_R \alpha, \qquad |\alpha| \le 1. \tag{9}$$

These conventions include empty revision support and zero exposure.

For binary Brier loss, fix probabilistic predictions a_c, b_c and binary targets. With $\Delta_c = a_c - b_c$ and $\delta_c = y'_c - y_c$, expanding the squared losses gives

$$h_c = -2\Delta_c\delta_c, \qquad E = 2\sum_{c\in\mathcal{R}} w_c|\Delta_c\delta_c| = 2\sum_{c\in\mathcal{R}} w_c|\Delta_c| =: B. \tag{10}$$

The signed equality is the finite weighted binary specialisation of the squared-loss pairwise shift in Kuang (2026, Proposition 2.1).

For separate bookkeeping, let $\mathcal{C}$ be the original support and consider pure removal of $\mathcal{A} \subseteq \mathcal{C}$ while targets, predictions, and the weighting rule stay fixed. Put $W_A = \sum_{c\in\mathcal{A}} w_c$ and $G_A = \sum_{c\in\mathcal{A}} w_c g_c$. For $W_A < 1$, renormalising the retained weights gives

$$d_{-\mathcal{A}} = \frac{d_v - G_A}{1 - W_A}, \qquad d_{-\mathcal{A}} - d_v = \frac{W_A d_v - G_A}{1 - W_A}. \tag{11}$$

Removal of all mass, $W_A = 1$, leaves an undefined evaluation. The changed denominator distinguishes removal from a target-value intervention.

Exact-rational validation covers 28,561 MAE cases, 676 binary-Brier cases, 10,000 weighted panels, 49 tie/reversal cases, seven factorisation cases, and 49 support-removal cases. A second implementation, written without sight of the first, reports no violation.

Proof of Proposition 3. Write $\delta_i = s_A(i) - s_B(i)$ and $S = \sum_i \delta_i$, so $d = S/N$ and $\sigma S > 0$. An admissible revision relabels or deletes items of R; the sign of the revised margin is that of the revised sum divided by the revised count. Relabelling item i to key g' changes $\sigma\delta_i$ to $\sigma\{\mathbf{1}[\hat{y}_A(i) = g'] - \mathbf{1}[\hat{y}_B(i) = g']\}$, which is -1 for some g' exactly when the two answers differ (choose the challenger's answer), and is otherwise unchanged; hence a challenger gains 2 on P_R, 1 on Z_R, and nothing elsewhere. Deleting item i subtracts $\sigma\delta_i$ from the sum and one from the count: on P_R this lowers the mean whenever the current mean is below one, on items with $\sigma\delta_i = -1$ it cannot lower it, and on neutral items it lowers the mean only when the current sum is already negative, in which case deleting all of Z_R^0 is optimal. When relabelling is allowed, changing a separable neutral item to -1 is at least as adverse as deleting it, since all revised means lie in $[-1, 1]$. For a fixed number k of deletions from P_R, the best revision therefore relabels every remaining item of $P_R \cup Z_R$ if

permitted, and deletes either none or all of Z_R^0. Under deletion alone Z_R is empty, so this step covers every neutral item. Empty supports are excluded. If the neutral set is nonempty and deleting all of it would leave an empty support, the remaining numerator is zero and retaining them attains the same minimum mean. This proves equation 7; the sign is preserved for every revision exactly when the minimum over k and j of the revised mean stays positive, that is when $E(R) < |d|$. The flip budget minimises the number of edits subject to the same accounting, using the cheapest edits first (relabels on P, then single-unit edits). Exact enumeration of every admissible revision on $3{,}000$ random instances checks both implemented formulas. A separate exact-arithmetic regression includes both margin orientations and all small outcome types, comparing $5{,}772$ cases with $100{,}876$ nonempty revisions and finding no displacement mismatch. It also includes an explicit deletion-only example: with contributions $(1, 1, -1, 0)$ and the two positive and the neutral items deletable, the largest displacement is $5/4$. Deleting the neutral item matters even when the models' wrong answers differ; all neutral items therefore belong to Z_R^0 in this mode.

C TWO OFFICIAL TRAFFIC ARCHIVES

The two Caltrans PeMS trees contain the same station–time identifiers on the paired days. Released TrafficStream targets match the current reprocessed tree. The other tree is a legacy official version; neither archive is designated as ground truth by this comparison. Predictions from the four released methods stay fixed while the key changes. Reconstructed or unmatched target cells are excluded from the primary paired support.

The four-year pooled MAE over the first three forecast steps, through 15 minutes, changes from 12.795 to 22.195. Eight of the 72 annual pairwise signs reverse, while all 18 pooled signs remain unchanged. The four-year criterion was registered before the legacy targets were read and retains its null outcome. The annual classification was specified after the first annual scores were known. The pooled ratios of the two methods' errors are 0.948, 0.951 and 1.003 at 15, 30 and 60 minutes. The certificate retains the shorter-horizon comparisons and abstains at the longest horizon.

An independent raw-input reconstruction matches all $7{,}384$ comparable result fields. An earlier score comparison reported failure because 41 of 312 rows differed in decimal serialization through trailing zeros; numeric values, signs and decisions agreed. That failure remains distinct from the subsequent successful raw-input reconstruction.

C.1 CORRECTIONS TO THE SURVEY IMPLEMENTATION

The original survey parser selected the quarter after the target even if that release did not contain the target. The corrected parser selects the first later vintage that does. It also aggregates monthly unemployment observations into their intended quarters. Independent rational arithmetic identifies exact zero margins that floating-point summation can perturb; the strict-order analysis treats margins below 10^{-9} as ties. The reported results use the corrected rules, and earlier outputs are preserved.

D DENOMINATORS FOR ABSTENTION STRATA

Table 4: Observed flips among abstained pairs in Table 1. Each entry gives the number of flips divided by the number of pairs in the exposure-ratio bin; an empty bin is shown as –. The fractions describe the same model-sharing comparisons as the main table.

Series	$[1, 2)$	$[2, 5)$	$[5, 10)$	$[10, 10^9)$
MMLU-Redux	0/118	0/64	0/28	7/22
GSM8K-Platinum	0/94	0/57	0/28	8/25
SWE-bench Verified	0/24	0/17	0/7	2/8
SWE-bench Lite	0/33	0/37	0/13	4/14
SPF GDP, h=0	0/13	12/246	32/258	123/302
SPF GDP, h=1	0/5	3/164	41/290	124/358
SPF unemp., h=0	0/224	4/265	22/110	41/98
SPF unemp., h=1	0/238	0/215	4/85	21/79
TIME Job Claims	0/79	0/55	3/26	0/16
TIME US Labor	0/57	0/37	1/7	3/7
TIME Housing	–	–	–	–
TIME Vehicles	0/41	0/34	0/6	3/23

E THE SURVEY SERIES UNDER A FIXED-LAG KEY

The one-year key produces 267 reversals and retains 501 comparisons, compared with 427 and 249 under the latest key. Table 5 holds the models and scored quarters fixed. This lag sensitivity was registered after exposure to the historical comparisons. Exact arithmetic excludes boundary ties $E = |d|$ from certification.

Table 5: Fixed-lag and latest survey keys on the same nonzero comparisons. The one-year key is the first vintage issued at least four quarters after the first release. Certification uses the strict realized-revision bound.

Series	Second key	Pairs	Certified	Reversals
GDP, $h = 0$	One year	819	1	85
GDP, $h = 0$	Latest	819	0	167
GDP, $h = 1$	One year	817	0	127
GDP, $h = 1$	Latest	817	0	168
Unemployment, $h = 0$	One year	781	197	36
Unemployment, $h = 0$	Latest	781	84	67
Unemployment, $h = 1$	One year	782	303	19
Unemployment, $h = 1$	Latest	782	165	25

E.1 TEMPORAL DEPENDENCE OF THE GDP NOISE REFERENCE

Calendar-block sensitivity retains dependence between adjacent quarters. The post-exposure analysis uses every registered block length $L \in \{1, 2, 4, 8\}$ and the original comparison set. For each model pair, let δ_t be its original-key loss difference on an observed quarter, d its observed mean and n its number of jointly observed quarters. On the full calendar of T quarters define $x_t = \mathbf{1}\{t \text{ observed}\}(\delta_t - d)$. Missing quarters have zero centered influence, preserving their original positions and the gaps between observations. Let $S_\ell(s) = \sum_{a=0}^{\ell-1} x_{(s+a) \bmod T}$ and let v_ℓ be the variance of $S_\ell(s)$ over uniformly distributed calendar starts, with $v_0 = 0$. Drawing independent circular blocks to fill exactly T positions, with the final block truncated if needed, gives

$$s_L^2 = \frac{\lfloor T/L \rfloor v_L + v_{T \bmod L}}{n^2}.$$

This is the exact variance of a centered-influence perturbation with a fixed denominator, conditional on the observed calendar and missingness. Its denominator remains fixed, and the interval describes those draws. At $L = 1$ it equals the exact empirical independent-item bootstrap variance.

For each L, the Gaussian reference has expected flip count $\sum_p \Phi(-|d_p|/s_{p,L})$, with zero contribution when $s_{p,L} = 0$ and $d_p \neq 0$. This analytical expectation refers to the Gaussian approximation; the block distribution can have different sign-flip probabilities. Linearity justifies summing dependent comparisons, while aggregate variances, confidence intervals and multiplicity corrections remain unestimated. The strict threshold remains $|d_p|/s_{p,L} > 1.6448536$.

Table 6: GDP noise-reference sensitivity on the unchanged $1{,}636$ comparisons, of which 335 reverse. The threshold column counts observed reversals beyond the declared sampling threshold. The last column takes the median after computing each pair's spread ratio.

Block length	Reference flips	Beyond threshold	Median U	Median s_L/s_1
1	305.53	21	0.555	1.000
2	303.72	17	0.568	1.000
4	294.64	15	0.580	0.970
8	282.83	19	0.611	0.935

The calendar covers 140 quarters; the median pair is observed on 42, with a minimum of 20. Observed gaps within each pair's span are preserved for every block length. An independent implementation parses both original workbooks and matches the $6{,}544$ pair-by-length rows. The sign reversals and deterministic revision envelopes are unchanged. Sparse calendars and the Gaussian approximation limit the interpretation of these noise-normalized counts.

F Retrospective intersection of sampling and revision rules

The replay retains all $8{,}535$ current pairs, including 14 old ties that cannot support a strict ordering. Among the $8{,}521$ eligible pairs, 458 reverse and three become ties. The Gaussian item reference uses $z > 1.6448536$. A nonzero margin with zero standard deviation has $z = \infty$; an old tie or unavailable spread cannot pass the interval rule. Zero spreads are retained exactly.

For categorical labels, orient each old item contribution q_i toward the old leader. Let $m_i \leq 0$ be its minimum under one common admissible label, with invalid predictions remaining incorrect, and set $g_i = q_i - m_i$. The smallest number of sorted gains whose sum reaches $|Nd|$ is the exact edit count for losing a strict sign, including a tie. Relabelling dominates deletion for this criterion: its numerator reduction is at least the deleted contribution, and it keeps the denominator positive. Thus a budget K preserves the sign precisely when this count exceeds K, or loss of the sign is unreachable. For deletion alone, the threshold is $|Nd|$ when $|Nd| < N$; a pair agreeing in direction on every item retains that direction on every nonempty subset. These elementary thresholds apply to each comparison separately; joint adversarial rankings require additional constraints.

The published-count scenario uses deletions plus retained label changes, without giving the rule their locations. The alternative rounds $N/20$ to the nearest integer, with half-integer ties rounded to an even integer. All $4{,}022$ eligible categorical pairs have actual revisions outside this alternative budget. Numeric budgets are the unchanged historical ninetieth-percentile envelopes. Their actual-containment fields are unknown, not false; they nominally retain $1{,}075$ pairs in either scenario. The supplemental pair table separates nominal retention, containment status and certification of the realised revision. Prospective calibration remains unestablished. These rows retain the relaxed bound $2\sum_i w_i \min\{\rho_i, |a_i - b_i|\}$; the separate exact-box replay compares it with the endpoint calculation. A source-level reconstruction matches the complete item support for all $4{,}502$ numeric pairs before reusing their cached budgets; three of them are tied under the original key and are excluded from the strict-order tables, which report $4{,}499$. Pointwise containment of a revision is a separate condition.

The replay was registered after historical outcomes were known. A separate implementation uses gain histograms and independently reconstructed source supports, matching the pair tables and summaries. It also checks the full support-admission predicate; the initial implementation's shorter predicate is restricted to this fully matched input. The counts describe comparisons sharing models and items; their dependence precludes an independent-trial confidence interval.

The published-count budget retains $1{,}049$ GSM8K, 750 MMLU, seven survey and $1{,}068$ TIME comparisons, totaling $2{,}874$. The retained SWE-bench count is zero. Its seven missed reversals and the sampling rule's 22 missed reversals are disjoint.

Table 7: Nominal retention in the aligned replay, conditional on the stated revision class. The five-percent scenario changes only the four categorical budgets; their actual revisions are outside the class. Numeric containment is unknown. Retention is not verified coverage of the realised revision, and neither outcome column is a future error rate.

Rule	Kept	Reversals	Revised ties
Item-sampling interval	5,712	22	0
Budget, published touched count	2,874	7	0
Budget, five-percent scenario	3,731	7	0
Both, published touched count	2,506	0	0
Both, five-percent scenario	3,363	0	0

G Data and computation index

The supplementary derived records cover every one of the twelve series in Table 1: aligned MMLU, aligned GSM8K, SWE-bench Verified and Lite, four survey series and four TIME tasks. They contain pair identities, old and new margins, item moments, revision bounds and strict decisions. Source comments identify the underlying numerical records without changing the rendered manuscript. The other derived tables supply matched subset distributions, shared-key controls, exact edit thresholds, judging-protocol contrasts and checkpoint decisions. Appendices M and N distinguish reconstruction from raw outputs from reconstruction using supplied comparison statistics.

H Registered checkpoints

The original registration froze the four TIME tasks on 2026-09-04 using the newest key already stored for each series. Historical displacement envelopes span eleven weekly or three monthly releases; the ninetieth percentile is primary, with the ninety-ninth percentile and historical maximum retained as secondary levels. The forecasts, cells, weights and all 1,300 pair decisions remain fixed.

The calendar extension registered on 2026-09-05 selects the latest actual series vintage at or before each declared date, regardless of execution time. Missing or nonfinite registered cells make the whole task incomplete, preserving the registered support and weights. The original later-release resolver uses a different missing-value rule, so any future output with dropped cells must be reported as incomplete for this fixed-support claim.

Table 8: The first registered checkpoint, as of 2026-09-14. Every task retains 325 pair decisions; the final columns give exact-box certifications at the registered historical percentile levels, computed on this checkpoint's own forward cells rather than on the historical test window, so they are not comparable with the replay counts of Appendix L. Certified comparisons have zero reversals.

Task	Cells	Changed	Reversals	p90	p99	Maximum
Job Claims	104	0	0	0	0	0
US Labor	240	0	0	277	223	67
Housing Inventory	144	62	1	259	0	0
Vehicle Supply	300	0	0	209	76	0

Retention is a steep function of the declared budget rather than of the data: on this checkpoint US Labor certifies 277, 223 and 67 comparisons at the ninetieth percentile, the ninety-ninth and the historical maximum, and Housing 259, zero and zero. The quantile is an input the analyst declares, not a quantity the method estimates, and the guarantee is conditional on it. The single reversal falls on a pair that abstains at every level. Housing has one of 144 cells outside its p90 envelope and none outside the larger envelopes. The other three tasks retain every registered cell value. Housing's changed values use a 2026-09-03 vintage, earlier than the registration, so this checkpoint is descriptive. An independent offline computation reconstructs all cells and decisions from the archived responses and frozen predictions. The primary added checkpoint is as of 2026-11-02; the original 2026-11-16 and 2027-02-16 resolutions remain prospective tests.

I HISTORICAL-VINTAGE SENSITIVITY

The provider of the TIME leaderboard targets retains 74 to 341 vintages per series, of which 8 to 42 cover every scored cell. Rescoring those complete vintages with predictions, cells and weights fixed gives the empirical spread of each margin across archived keys. The registered question was whether that spread reaches half the item-sampling spread. The median ratios are 0.09 and below 0.04 on the computable tasks, below that threshold. These results concern the retained historical vintages and their small original margins.

A split-conformal displacement quantile (Vovk et al., 2005), calibrated on vintages before the median release date, retains 313 of 325 claims-task pairs at $\alpha = 0.10$, with no reversal among those 313 retained comparisons on the later held-out vintages. Three comparisons reverse in the full set of 325. The conditional guarantee assumes that later and calibration vintages are exchangeable. Seasonal revisions can violate this assumption, and the other tasks have too few calibration vintages. This historical split is separate from the registered future-release test in Appendix H.

The fixed-lag survey analysis appears in Appendix E, the intersection of sampling and revision rules in Appendix F, and historical budget failures in Appendix L.

What the held-out conformal result does and does not show. Specified after the result was known, this descriptive analysis applies to the conformal rule the matched-retention test used for the budget rule. At $\alpha = 0.10$ the rule certifies 313 of the 325 claims-task comparisons and none reverses on the 21 later vintages; three of the 325 reverse in total. Held to that retention, ranking by the original margin over the item spread reverses two. That is the whole of the apparent gain, and it does not come from the conformal construction. Ranking by the original margin over the standard deviation of the same calibration vintages selects exactly the same 313 comparisons, verified by array equality, and also reverses none; discarding the comparisons whose sign already changed during calibration retains 322, nine more, with none. Under the item-spread ranker the largest zero-reversal retention is 291, not the 245 the registered threshold happens to select.

Why the historical threshold retains one more pair. The current main analysis uses Equation 1, an analytic population plug-in spread, and retains 244 claims comparisons at the same threshold. The historical conformal branch estimates that spread from $1{,}000$ iid item-bootstrap replicates and retains 245. The support is the same 104 ordered items; original margins agree to floating precision, with a maximum absolute difference of 1.78×10^{-15}. The two spread estimators target the same conditional bootstrap variance, but a finite simulation need not equal the analytic value. This is not an item-divisor substitution. For TimesFM-3 versus Toto-2.0-313m, the historical and analytic spreads are 0.0756444 and 0.0763162. Their standardized margins, 1.64656 and 1.63207, straddle the unchanged threshold 1.6448536. This is the sole retention change. The supplement lists every pair under both contracts and preserves the historical result; the original bootstrap draws were not archived and are not recreated by this analytic reconciliation. The same-calibration-information comparison above remains unchanged.

Coverage limitations. Three facts bound the reading further. The 42 vintages carry only three distinct answer-key states, of 15, 2 and 25 vintages, so the quantile rests on two revision events and not on 42 exchangeable draws; the study's own stationarity diagnostic is 1.06×10^{10}, which rejects exchangeability, so no coverage guarantee is in force; and moving the split so that the later revision falls on the test side certifies all 325 comparisons and reverses all three, the worst outcome of any rule we tried. The other three tasks have four to five calibration vintages, at which the quantile at this level is infinite and nothing is certified, so the analysis does not extend.

J ARCHIVED JUDGING PROTOCOLS AND SAMPLE FORMATION

Source and intervention. Published AlpacaEval per-instruction annotations (Li et al., 2023) provide every candidate output and judgment in this analysis. File hashes and byte counts identify the downloaded corpus. The downloader used the upstream main branch. Individual file hashes identify the downloaded observations; the inventory commit identifies the earlier listing.

The protocol contrast compares `weighted_alpaca_eval_gpt4_turbo` with `alpaca_eval_cot_gpt4_turbo_fn`. Both configure `gpt-4-1106-preview` and the same `gpt4_1106_preview` reference outputs. System and reference output strings agree on the common instructions. Prompt, response format, decoding settings and parser change: a one-token log-probability preference is compared with a function-call ranking that requests an explanation. The stored judge completions also differ. The comparison concerns a complete scoring protocol; Proposition 3 instead assumes a shared categorical label.

The joint-change contrast compares `alpaca_eval_gpt4` with `weighted_alpaca_eval_gpt4_turbo`. The reference changes from `text_davinci_003` to `gpt4_1106_preview`; older records include both `alpaca_eval_gpt4` and `alpaca_eval_gpt4_0613` annotator identifiers. This changes judge protocol/model and reference jointly, so the effect includes all of these changes.

Selection and its sensitivity. The protocol archive contains thirteen available systems. We exclude `gemini-pro` because stored outputs differ on 66 of 805 common instructions. The initial analysis also excluded `gpt4_1106_preview` because its score vectors were identical. This was an outcome-based exclusion: that reference self-comparison is legitimate even with unchanged scores. A separately registered post-exposure sensitivity includes it without changing other rules. We report that complete fixed-output intersection in the main text and retain both versions in Table 9. The joint-change archive contains 51 systems; excluding `falcon-7b-instruct`, whose outputs differ on 800 of 805 common instructions, leaves 50. Systems with unchanged score vectors remain included. Model identifiers, exclusions and pair-specific support hashes accompany the released comparison tables.

Table 9: Sensitivity to a legal zero-change reference system. Pairs exclude original-margin ties. The protocol analysis retains fixed system and reference outputs. The joint contrast also changes the reference. All certification counts use the realised change.

Analysis	Systems	Pairs	Median U	Certified	Reversed
Protocol, initial filtered set	11	55	3.152	3	7
Protocol, all fixed outputs	12	66	2.910	13	7
Judge and reference jointly	50	1,222	5.292	87	115

Scoring and denominators. A finite preference in $[1, 2]$ maps to $s = \text{preference} - 1$; null and out-of-range values are excluded without imputation. There are 54 nulls and 188 zero-valued sentinels in the corpus, with the zero sentinels in identical-output records under the older configuration. Each pair uses the four-way intersection of valid instructions for both systems under both protocols, uniform weights, and at least 100 instructions. Support is fixed across the two scores for a given pair but can differ across pairs: 804 to 805 instructions for the protocol contrast, 757 to 805 for the joint change. The latter has 1,225 total pairs, including three original-margin ties; the reported denominator is therefore 1,222. Each reported ordering concerns its own paired support; a whole-leaderboard ranking would require a common-support design.

Let $c_i = s_A(i) - s_B(i)$ and $c'_i = s'_A(i) - s'_B(i)$. We compute

$$d = \frac{1}{n}\sum_i c_i, \qquad d' = \frac{1}{n}\sum_i c'_i, \qquad E_J = \frac{1}{n}\sum_i |c'_i - c_i|, \qquad |d' - d| \le E_J. \tag{12}$$

This generic triangle bound applies to the realized scoring change. The MAE projection and shared-label formulas impose their additional loss-specific structure. Original margins within 10^{-9} of zero are ties. The item spread is

$$\widehat{\mathrm{se}}(d) = \sqrt{\left\{ n^{-1}\sum_i c_i^2 - (n^{-1}\sum_i c_i)^2 \right\}/n}. \tag{13}$$

The original plan specified a bootstrap for nonbinary scores; both implementations used this analytic spread, a disclosed deviation. $U = |d' - d|/\widehat{\mathrm{se}}(d)$ expresses displacement in item-sampling units;

simultaneous and temporal inference require additional assumptions. Certification uses $E_J < |d|$; the descriptive interval rule uses $|d|/\widehat{\mathrm{se}}(d) > 1.645$.

The protocol contrast retains 58 comparisons under the interval rule, four of which reverse; the joint change retains 1,091, with 60 reversals. The original protocol subset gave a median of 25 instructions to reach a tie; including the reference changes this to 36.5. For the joint change the median is 46. These edit budgets and realized certificates describe the archived changes; the next judge upgrade remains unobserved. The original extraction-only contrast is excluded because its upstream annotator configurations are mixed. An independent implementation reproduces the zero-change sensitivity from raw archived annotations; the original strict comparison table also agrees with the corrected scoring run.

K A CONTEMPORARY ARCHIVED JUDGE CONTRAST

Population and exposure. The Arena-Hard archive (Li et al., 2024) contains matched answer files and both Gemini and GPT-4.1 judgments for 27 candidate systems. We include all such systems, including the judge models' own answers. The downloaded corpus has 84 byte-verified files: candidate answers and both judgments, the two task-specific reference answers, and the question file. The two references differ, so hard prompts and creative writing are separate populations, with task-specific scores. The analysis plan preceded full-corpus effect calculations but followed exposure to the published aggregate table and two schema samples. Every answer and judgment comes from the existing archive.

Raw contrast. We orient the contrast from Gemini to GPT-4.1. Both score the same archived candidate and reference outputs, while their judge identities and decoding configurations change together. Each question averages the two answer-order games, with a candidate win, tie and loss scored as 1, $1/2$ and 0; strong and slight preferences have equal weight here. The official metric additionally weights preference strength and adjusts for style. Pair-specific comparisons require valid inputs and scores for both candidates under both scorers. We report $U = |d' - d|/(\widehat{\mathrm{sd}}(c)/\sqrt{n})$, using the sample standard deviation with denominator $n - 1$. This is the one place the paper uses $n - 1$: the item-sampling unit of Equation 1 and its appendix forms use the plug-in n divisor, because there the spread is a population quantity of the scored items; here the contrast is a per-question sample of judged games. Whole-roster winners instead use the common valid questions across every candidate and scorer. Missing or invalid games receive intervals $[0, 1]$ for the separate fixed-population analysis.

Table 10: Raw archived contrast. N is the fixed question count and $n_\cap$ the whole-roster common support. Reversals use pair-specific support; robust reversals persist under all missing-game completions on the fixed population. Regret is the original winner's gap to the alternate scorer's winner on common support, in percentage points.

Task	N	$n_\cap$	Median U	Reversals	Robust	Regret
Hard prompts	500	494	2.340	31/351	29	9.767
Creative writing	250	245	2.238	30/351	30	2.755

Under both raw task scores, Gemini leads under its own judgments and `o3-2025-04-16` under GPT-4.1 judgments. That pattern is what the self-preference literature describes (Panickssery et al., 2024); with one judge pair and one archive we report it and do not separate it from the other protocol differences that change together. These winners also remain strictly ahead under the fixed-population missing-score intervals. Shared-question bootstrap draws preserve dependence among candidates: within each task, all comparisons use the same resampled questions in each of 2,000 replicates. The percentile intervals describe item sensitivity conditional on the archived scorers; future judges and human quality remain outside this analysis.

Empty-output admission contracts. One archived Gemini answer is an empty string. The primary implementation treats it as unavailable input; the independent implementation admits the observed empty output. The plan did not resolve this distinction. Their comparison retains 234 field

differences, affecting 25 pair-specific point/support rows and 26 pair-specific interval rows. A separate missing judgment for Gemini Flash excludes the same question from whole-roster common support. The independent reconstruction agrees on the table's category aggregates, both common-support winners and regrets, the fixed-population winner sets, and the registered bootstrap results. The complete pair tables retain the stated discrepancies, and upstream corpus completeness remains unverified. Both original executions are preserved. The distinction is now traced to the input test: the primary code requires a nonempty answer, whereas the other accepts any string other than a missing value. The empty string passes its substring check and is admitted with the recorded zero score under both judges. The primary analysis continues to use the nonempty-answer contract: the archived empty answer is unavailable for paired point estimates and bounded in $[0, 1]$ for the fixed-population analysis. This is an explicit post-exposure description of the existing choice, not a new preregistration. A per-field table reconciles all 234 differences from the saved arrays; it does not repair the original completeness failure or claim agreement of every pair table. Future studies must fix this admission rule before observing effects.

The official metric changes the interpretation. We also execute the author's archived scoring functions. Strong preferences contribute three battle rows; other preferences contribute one. A question with either game missing is excluded from that judge's native score. For hard prompts, the published protocol fits a Bradley–Terry model with length and markdown controls and reports the median over 100 bootstrap fits. Creative writing uses the mean over 100 strength-weighted bootstrap scores, with the recommended combined configuration concatenating both judges. These native populations and resampling units differ from our paired, equal-question analysis. The optimizer returns finite values; global optimality remains unverified. The historical random seed is unspecified, so numerical comparisons use our declared fixed-seed execution.

With the official hard-task style controls, `o3-2025-04-16` wins under both judges: the official hard-task champion remains unchanged. For creative writing, Gemini wins under Gemini judgments and under the combined configuration; o3 wins under GPT-4.1 judgments. Champion comparisons therefore depend on the scoring protocol. A dynamic-library error prevented the initial hard-task fits; an environment-only repair executed those failed cases without rerunning the successful creative-writing cases. The original failure and both runs remain in the artifact. One independent reconstruction of all configurations agrees on the displayed scores, intervals and ranks. It rebuilds the raw records, design matrices and resampling, but shares the unchanged author optimizer, fixing the numerical optimization component across implementations.

L LEGAL-KEY CONTROLS AND EXACT EDIT THRESHOLDS

L.1 ALIGNED SCORING AND A MATCHED RANDOM CONTROL

This analysis uses the current aligned MMLU and GSM8K panels. Each model's predictions are fixed, and the same parser scores the original and revised keys. The primary MMLU arm requires exact prompt identity; whitespace matching is a separate sensitivity arm. Let S denote the original support and T the published retained support. We distinguish deletion, comparing the old key on S and T; relabelling, comparing the two keys on T; and their joint effect, comparing the old key on S with the revised key on T. The deletion count $D = |S \setminus T|$ and the number C of changed canonical keys on T are operational edit counts. Metadata flags with unchanged parsed keys contribute zero relabellings. Table 11 reports these distinctions.

The control uniformly deletes D items for the deletion contrast, uniformly chooses C items from T for relabelling, and deletes D items from S before relabelling C distinct survivors for the joint contrast. Every changed key is shared by all models answering that item. MMLU draws uniformly from the other legal answer letters. GSM8K draws uniformly from the distinct nonmissing canonical predictions and an additional, unpredicted answer class, excluding the old key. This last class represents a present legal answer unmatched by every frozen prediction. Sampling is uniform over these finite score-equivalence classes; numerical answers and human correction mechanisms have different, unspecified distributions. Missing predictions remain incorrect.

We retain all 200 draws for each contrast, using seed 20260904 and the registered PCG64 stream. These choices preceded the new draws but followed exposure to historical results. Every registered model, contrast, seed and draw is retained. For each pair, $U = |\Delta d|/s$ uses the contrast-specific

Table 11: Aligned populations and observed edits. R is the metadata-marked set; No-op counts retained flagged items whose canonical key is unchanged. The two MMLU arms use overlapping items and models.

Panel	Models	$\lvert S\rvert$	$\lvert T\rvert$	D	C	$\lvert R\rvert$	No-op
MMLU exact	59	5381	5127	254	99	354	1
MMLU WS	59	5651	5388	263	100	364	1
GSM8K	60	1319	1209	110	10	120	0

baseline support (S for deletion and joint edits, T for relabelling) to compute the independent-item mean standard deviation, where $s^2 = [N\sum_t c_t^2 - (\sum_t c_t)^2]/N^3$. Original ties are retained in the data but excluded from nonzero-margin summaries and strict reversals. A zero standard deviation gives $U = 0$ for zero displacement and $U = +\infty$ otherwise; neither case is regularized by an artificial positive denominator. The tables report linear descriptive quantiles across the fixed random draws. A separate implementation replayed the registered stream and reproduced the edit plans, per-model scores and pairwise summaries. This is computational verification of the same randomization contract using the same draws.

The matched-control table for the single operations is Table 2 in Section 4. The joint-edit rows, which apply both operations at their observed counts, are below.

Table 12: Matched random controls for joint edits, columns as in Table 2.

Panel	Action	D	C	Flips	Random	U_{50}	Random
MMLU exact	joint	254	99	7	6.00 [3.00, 10.00]	0.287	0.490 [0.361, 0.616]
MMLU WS	joint	263	100	5	5.00 [2.00, 9.00]	0.305	0.477 [0.359, 0.611]
GSM8K	joint	110	10	8	7.00 [3.00, 13.00]	0.470	0.267 [0.189, 0.345]

Deletion has larger median displacement than the control's upper quantile in all three panels. Relabelling alone has smaller median displacement than the control's median in every panel. The joint MMLU median also lies below its control median, although its share with $U \geq 1$ exceeds the control's upper quantile. GSM8K's joint displacement is larger by both summaries. None of the joint strict-reversal counts exceeds the corresponding upper control quantile. These mixed results support separating deletion from key changes. The controls condition on observed edit counts and answer classes. Future revision risk, label correctness and model-selection performance remain outside these conditional comparisons.

L.2 The minimum number of legal edits

For an originally strict comparison, orient the pair so that its integer score difference is $d = \sum_t c_t > 0$, with $c_t \in \{-1, 0, 1\}$. Let $\ell_t = \min_{y \in \mathcal{Y}_t} c_t(y)$ over legal shared keys and $g_t = c_t - \ell_t$. For an edit with positive gain, a minimizing key necessarily differs from the old key. Sorting these gains gives the exact thresholds

$$k_\circ = \min\left\{k : d - \sum_{r=1}^{k} g_{(r)} \ \circ\ 0\right\}, \qquad \circ \in \{\leq, <\}. \tag{14}$$

An empty feasible set has no finite threshold. Thus $k_\leq$ is the number of edits needed to lose strict sign preservation, including a tie; $k_<$ requires a strict reversal. For an original tie we record $k_\leq = 0$ and leave $k_<$ undefined. A budget smaller than $k_<$ rules out reversal while allowing a tie when it reaches $k_\leq$.

The formula follows by separability: any set of k relabelled items can reduce the numerator by at most the sum of their gains, and choosing the largest gains attains that bound with legal shared keys. In these domains g_t is zero, one or two, so the threshold can also be obtained from the counts of the positive gains. For pure deletion, a useful deletion removes a positive contribution. Its thresholds are d and $d+1$, respectively, provided that many positive-contribution items exist and fewer than all support items are deleted. Otherwise the threshold is unattainable.

Table 13: The upper tail of the same control: the percentage of originally nonzero pairs displaced by at least one sampling standard deviation. All arms and both the median and upper-tail endpoints are retained.

Panel	Action	Observed (%)	Random median [p05, p95] (%)
MMLU exact	delete	9.3	0.0 [0.0, 0.0]
MMLU exact	relabel	1.6	21.2 [13.1, 25.8]
MMLU exact	joint	28.2	22.2 [12.3, 27.6]
MMLU WS	delete	8.1	0.0 [0.0, 0.0]
MMLU WS	relabel	0.9	20.0 [12.0, 25.8]
MMLU WS	joint	28.4	20.6 [10.7, 28.0]
GSM8K	delete	24.0	0.0 [0.0, 0.5]
GSM8K	relabel	0.0	0.0 [0.0, 0.0]
GSM8K	joint	26.2	0.4 [0.0, 3.7]

Unrestricted classification reduces to a margin threshold. Suppose scores are uniformly weighted classification accuracies, both models' predictions are legal labels, every item is editable, every legal replacement label is allowed, and each edited item costs one. Write the old leader's integer advantage as $H = N|d| = P - Q > 0$, where P counts items on which only the leader is correct and Q counts the reverse. Then

$$k_{\leq} = \lceil H/2 \rceil, \qquad k_{<} = \lfloor H/2 \rfloor + 1, \qquad \text{strict preservation at budget } K \iff H > 2K. \quad (15)$$

Each edit reduces the advantage by at most two. Relabelling a leader-only correct item to the challenger's answer attains that reduction, and $P \geq H$ supplies enough such items for either threshold. This proves the statement, including $H = 1$. Here $k_{\leq}$ reaches a nonpositive margin, not necessarily an exact tie: an odd binary-classification advantage can reverse without ever tying. This is an elementary specialization of Equation 14, not an additional algorithmic contribution. It provides no ranking information beyond the old margin. Restricted editable sets, invalid predictions, unequal weights or costs, and coupled label constraints need not obey this simplification.

For a joint edit on S, deletion and relabelling each cost one action, and an item cannot receive both. Here each pair admits a legal key with zero contribution: MMLU has more answer letters than the pair can predict, and GSM8K includes an unpredicted answer class. Consequently $\ell_t \leq 0$, and relabelling can reduce the numerator by at least as much as deleting that item. Replacing useful deletions by these relabellings preserves the support and cannot increase the numerator. The joint sign threshold therefore equals the relabelling threshold on S; the separately reported relabelling contrast is instead on T. This argument is specific to the declared additive score and legal domains. Weighted costs, coupled annotation constraints and other metrics require their own feasible-set calculation.

All nonzero MMLU comparisons admit both endpoints. GSM8K has 238 deletion-only comparisons with no strict reversal; the corresponding counts for relabelling on T and joint edits on S are 112 and 112. In contrast, their tie endpoints are attainable. Irreversibility can reflect missing-output and agreement structure, rather than a model's semantic robustness. For example, a model with invalid outputs on every item has zero accuracy under every shared key. Separate finite subsets explain why a strict-reversal median in the table can be smaller than the tie-threshold median.

The joint budget permits any mixture of deletion and relabelling at any item, including mixtures and locations different from those observed. It is therefore distinct from the realised-revision bounds elsewhere in the paper. A comparison outside the protected set has a legal counterexample within the budget, whether or not the actual revision attains that counterexample. All pair endpoints and emitted counterexamples were checked against a second implementation using finite-action tests and direct witness scoring.

Minimum-edit robustness is an established idea, including training-label poisoning (Mohgaonkar et al., 2026). The calculation here holds model outputs fixed, whereas training-label poisoning changes the fitted predictor. Contemporary reannotation studies also separate corrected labels, excluded cases and, in some branches, retraining (Volkov et al., 2026; Larson et al., 2026). We measure the resulting fixed-output comparison under explicit legal edit sets.

Table 14: Exact threshold summaries on all nonzero comparisons. B uses the observed D, C, or $D + C$ for deletion, relabelling, or joint edits, respectively. Strict kept counts $k_{\leq} > B$; No reversal counts $k_{<} > B$, including unattainable endpoints. The medians use each endpoint's finite subset separately, allowing either ordering. The budgets use the observed historical edit counts.

Panel	Action	B	Nonzero	$k_{\leq,50}$	$k_{<,50}$	Strict kept	No reversal
MMLU exact	deletion	254	1710	569	570	1245	1245
MMLU exact	relabel	99	1708	271.5	272	1310	1310
MMLU exact	joint	353	1710	285	285	750	750
MMLU WS	deletion	263	1708	597	598	1251	1255
MMLU WS	relabel	100	1709	286	286	1312	1315
MMLU WS	joint	363	1708	298.5	299.5	758	758
GSM8K	deletion	110	1762	295	260	1376	1408
GSM8K	relabel	10	1761	152	142	1655	1665
GSM8K	joint	120	1762	160.5	150	1049	1065

L.3 Why a historical revision budget can fail

Future revision size requires a budget choice separate from the exact threshold. Under absolute loss a budget must be estimated or declared. Taking the ninetieth percentile of preceding historical revisions as the per-cell budget certifies $1{,}068$ of $1{,}300$ TIME comparisons, of which seven later reversed, and seven of $3{,}199$ survey comparisons, none of which reversed. The estimate misses seven of the ten TIME reversals. TIME uses the 60 observations before its test window; the survey uses up to 20 preceding quarters for each target, requiring at least 8. Raising the historical quantile to the largest preceding revision leaves those same seven failures because the relevant histories were nearly unrevised. Historical quietness therefore leaves future revision size uncertain. Sweeping the level from the median to the largest preceding revision reduces survey coverage from 10.0 per cent, 320 of $3{,}199$ comparisons, to zero, without a certified reversal at any level; the seven comparisons above are the ninetieth-percentile figure.

At the ninetieth-percentile budget, the exact box rule of Proposition 2 certifies 275 claims pairs against 93 for the relaxed bound, and 110 and 312 unemployment pairs against 1 and 6, with no reversal among the added pairs. This historical replay was registered after exposure and runs on the historical test window; the first registered calendar checkpoint, reported in Appendix H, runs on later forward cells, so its certification counts are a different quantity. Historical estimates and declared budgets have different interpretations.

Threshold sensitivities, not prospective rules. The sampling rule uses $|d|/s > 1.6448536$. At the two-sided value 1.959964 it retains $5{,}301$ comparisons with nine observed reversals; its conjunction with the budget rule retains $2{,}453$ with none. The conjunction's smallest retained $|d|/s$ is 1.646. Ranking by $|d|/s$ alone, the largest retention with no observed reversal under one strict threshold is $4{,}108$ of $8{,}521$ over all series and $4{,}977$ of $6{,}335$ on the eight series kept at all by the budget rule. Thresholds cannot split tied standardized margins. Both frontiers use observed outcomes and cannot set a threshold for an unseen revision. Categorical budgets disclose the observed edit count but not the edited locations, which is post-revision information in this replay.

Graded ratios and source stratification. Of $4{,}499$ comparisons with a numeric budget, the sampling rule retains $1{,}922$. None of the $1{,}019$ whose bound is at most twice the margin reverses, while nineteen of the 227 between five and ten times it do. All twenty-one retained reversals, however, fall in the two real-GDP series, which contribute one and zero comparisons below twice the margin. Permuting within series puts the observation at the 95th percentile of its null, against zero pooled. The apparent ordering is therefore between series rather than evidence of a useful within-series risk score. These are post-exposure diagnostics on dependent model comparisons.

M A common scorer for the two GSM8K keys

We retain the archived outputs and the full model roster, and score both keys with the same declared float/string parser. The original key is pinned to the dataset provider's released test file. The parser

Table 15: The conjunction against a sampling-only rule held to the same retention, on the same 8,521 comparisons. Matched retention selects the largest standardised margins, pooled or within each series. The retained identities differ: the conjunction shares 1,782 pairs with the pooled selection and 2,439 with the series-matched selection. Section 5 states the retention and reversal counts in the text.

Rule	Retained	Reversed
Sampling interval	5,712	22
Revision budget	2,874	7
Both rules	2,506	0
Sampling only, matched retention, pooled	2,506	0
Sampling only, matched retention, by series	2,506	0
Sampling only, on the eight series the budget rule keeps	4,757	0

Table 16: Sampling and budget rules on the same nonzero-margin comparisons, series by series. Their conjunction retains fewer comparisons and misses no observed reversal. Categorical budgets use the published edit count; numerical budgets use historical revision envelopes with unverified containment. Budget retention is nominal, not verified coverage of the realised revision. Section 5 states the totals and reads them.

		Sampling interval		Revision budget		Both rules	
Series	Pairs	Kept	Reversed	Kept	Reversed	Kept	Reversed
MMLU-Redux	1,710	1,600	0	750	0	750	0
GSM8K-Platinum	1,762	1,656	0	1,049	0	1,049	0
SWE-bench Verified	275	267	0	0	0	0	0
SWE-bench Lite	275	267	1	0	0	0	0
SPF GDP h=0	819	264	13	0	0	0	0
SPF GDP h=1	817	157	8	0	0	0	0
SPF unemp. h=0	781	341	0	1	0	1	0
SPF unemp. h=1	782	309	0	6	0	6	0
TIME Job Claims	325	244	0	93	0	93	0
TIME US Labor	325	216	0	325	4	216	0
TIME Housing	325	175	0	325	0	175	0
TIME Vehicles	325	216	0	325	3	216	0
Total	8,521	5,712	22	2,874	7	2,506	0

extracts the last final-answer marker; missing parses score zero under every key, and a missing gold answer fails the run. This defines the operational score used in both arms. The design was registered after the historical results were known, before the new joins and scores were computed. All retained Platinum keys are read, including unmarked items.

The registered endpoints distinguish archived correctness, the original key on all items, the original key on retained items, and the revised key on retained items. A further key-only counterfactual keeps all original items, using their original keys when Platinum omits them. The table reports strict sign changes, excluding a tie at either endpoint; these overlapping contrast counts describe separate scoring operations.

Contrast	Reversed pairs
Archived correctness to common-scorer original key	123
Original to revised key, full-support counterfactual	1
Original full support to original retained support	7
Original to revised key, retained support	1
Original full support to revised retained support	8

There are 2,699 changed model–item scores between archived correctness and the common-scorer original key. Appendix R.2 reconstructs a historical contract that reproduces every archived flag; its actual historical dependency revision remains unidentified. Under the common scorer, 10 retained

gold answers change, all within the declared editable set of 110 deletions and 10 metadata-marked corrections; scores outside that set remain unchanged. Missing parses remain in the population.

For this scorer we compute an exact conditional envelope directly. Orient each pair toward its original leader, write $\delta_i(y)$ for its score difference under key y, and let $K \subseteq R$ be the retained editable items. With uneditable items fixed, the smallest admissible oriented margin is

$$m_* = \min_{K \subseteq R,\, N-|R|+|K|>0} \frac{\sum_{i \notin R} \sigma\delta_i(y_i) + \sum_{i \in K} \min_{y \in \mathcal{Y}_i} \sigma\delta_i(y)}{N - |R| + |K|}, \qquad E = \max\{0, |d| - m_*\}.$$

The per-item key set contains each valid parsed model answer and another nonmissing token. Unparseable model answers remain incorrect, so they cannot become correct by choosing a missing key. A retained-count dynamic program evaluates this finite minimum; an independent implementation sorts the per-item minima for each retained count. This specialises the declared scoring contract, including its treatment of missing literal answers. Proposition 3 assumes categorical attainability instead.

On the 60 models, 1,319 original items and 1,209 retained items, this contrast has 1,762 nonzero original margins and eight ties, retained separately. It certifies 1,558 comparisons; none of the eight strict reversals is certified. The actual target scores lie inside the declared set, which is checked separately from evaluating the conditional bound. An independent raw-file implementation agrees on 1,346,775 fields. Its comparison first failed on an optional byte-count field for the added source; that schema-only repair is recorded without rerunning either raw calculation. The main table, legal-key controls and budget comparisons all use this aligned contrast.

N ALIGNED MMLU KEYS AND A COMMON SAMPLING REFERENCE

Population and identity. We keep all 59 models in the archived roster and all 57 subjects. The target inventory contains 5,700 annotations. The primary identity rule requires exact equality of the complete rendered prompt, including the question, ordered answer choices and answer marker, within its subject. A match must occur exactly once in the target inventory and in every model's source. A secondary analysis collapses Unicode whitespace while leaving other characters unchanged. Both rules preceded scoring. Whitespace-equivalent prompts can differ at the token level, so the secondary result is reported separately. Predictions are the maximal entries of the four-option output vectors, with the lowest option index breaking ties. Invalid predictions would score zero under every key without changing the support; none occurs on either calculable support.

The original keys are the original-label fields declared by Redux. The target parser uses the registered numeric and A to D letter mapping, with the documented option order; the standalone-letter interpretation is an explicit analysis assumption. Unknown targets remain distinct from exclusions. The calculable original set S requires a one-to-one identity, a legal original key and a known action. Its retained subset T contains usable revised labels. The other actions in S are exclusions under the registered annotation-status policy; the source annotations themselves give no item-deletion instruction. In the primary analysis, 289 annotations lack a common one-to-one source identity, 26 have nonunique target identities, and four otherwise matched annotations have unknown actions. These records remain in the inventory but outside the conditional comparison. Earlier annotation-status counts used a different join and treated metadata flags as edits. The present counts require common prompt identity and a changed canonical retained key; the primary arm has 254 exclusions and 99 retained key changes, while its 354 flagged items include one unchanged key.

Paired comparisons. The main contrast scores the original key on S and the revised key on T. A fixed-support contrast scores both keys on T. A further counterfactual keeps S, replacing keys on T but leaving the original keys on excluded items. The model outputs and argmax scorer are unchanged. All pairs, ties and subjects are retained in the corresponding records.

Conditional envelope. The editable set R is defined by the non-`ok` metadata statuses within S, before comparing model scores. Each may be removed or assigned one of the four legal keys; outside R, the original configuration is fixed. For each pair and retained editable-item count, we sum the smallest attainable item contributions and divide by the resulting positive support size. Minimising over these counts gives the worst oriented margin, as in Appendix M, using precisely the four legal

Table 17: Aligned MMLU results. Exact prompt equality defines the main analysis; whitespace folding is a separate sensitivity. Reversal counts exclude ties at either endpoint. Conditional certifications refer to the main S-to-T contrast and the edit family described below.

Quantity	Exact prompts	Whitespace sensitivity
Calculable original items $\|S\|$	5,381	5,651
Retained items $\|T\|$	5,127	5,388
Analysis exclusions in S	254	263
Metadata-marked editable items $\|R\|$	354	364
Nonzero original margins	1,710	1,708
Main contrast: reversed pairs	7	5
Fixed T, changed keys: reversed pairs	6	8
Counterfactual fixed S: reversed pairs	5	4
Conditionally certified pairs	1,478	1,484
Reversals within certified pairs	0	0
Subjects with positive original-winner regret, main contrast	6	7
Subjects with positive original-winner regret, fixed T	2	2
Global original-winner regret, main contrast	0	0

MMLU labels. One implementation sorts item contributions; an independent implementation counts attainable negative and positive contributions. Both also calculate the largest margin and retain attaining configurations. The observed targets stay inside the declared family, checked separately from the numerical inequalities. The bounds condition on this admitted inventory and declared edit family; excluded records and future edit families require separate assessment.

Selection consequences. The global original-key winner, Yi-1.5-34B-32K, remains a revised-key winner under both identity rules. On the primary subject-level main contrast, the affected subjects are abstract algebra, business ethics, logical fallacies, miscellaneous, professional law and public relations. Restricting both keys to T leaves positive original-winner regret on logical fallacies and public relations. The primary main-contrast regrets include $2/41$ in professional law and $1/47$ in abstract algebra. These fixed-benchmark score gaps describe the archived models; capability differences beyond these data remain unestimated. Every subject is retained, and the global winner remains unchanged under both released keys.

One sampling-reference definition. For each nonzero original margin, let x_i be its original-key per-item score or loss difference. The current twelve-series descriptive summary uses the analytical empirical independent-item bootstrap variance

$$s^2 = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad z = \frac{|d|}{s}, \qquad U = \frac{|d' - d|}{s}. \tag{16}$$

This summary and the aligned MMLU tables retain zero variances and use a denominator floor of 10^{-300}. Replay and legal-key control diagnostics retain their separately specified zero-spread rules. The aligned binary scores permit exact rational variance calculation. The real-valued series use the same closed form evaluated in floating-point arithmetic. GDP reuses the checked independent-item result from Appendix E.1; the remaining survey and TIME series use their frozen scored supports and scales. This variance is computed analytically.

The current twelve-series table has $8{,}521$ nonzero comparisons and 458 reversals. Summing the Gaussian tails gives 694.04 expected reversals; 391 observed reversals have $z < 1$, and 22 have $z > 1.6448536$, comprising 21 GDP comparisons and one SWE-bench Lite comparison. For MMLU, the primary median U is 0.287 for the main contrast and 0.220 when both keys use T; all primary reversals in both contrasts have $z < 1$. Linearity justifies summing expectations for these dependent comparisons. This descriptive Gaussian reference leaves the actual non-Gaussian reversal probability, aggregate interval and multiplicity adjustment unestimated. The matched controls and budget comparisons use this aligned population with their declared variance conventions.

Conditional resampling of corrected releases. We retain the original model rosters, aligned item identities, exclusion policy and nonzero-original-margin pair set. Each of $2{,}000$ draws assigns the

same item weights to every model and to both keys. The primary design samples with replacement within each MMLU subject and across GSM8K items. A second design resamples entire MMLU subjects, retaining every item in a selected subject, and reuses the corresponding GSM8K draw. We recompute both retained supports, the margins and Equation 16 for every draw. Sampled ties remain in the fixed pair set. Neither models nor the two correction events are treated as independent population samples.

Why the archived reading is the smaller one. U is a ratio, so the gap between the archived median and the resampled one is in its numerator, its denominator, or both. Specified after the gap was known, this descriptive analysis re-reads the frozen draws and fits nothing. Per design, the median $\widehat{\mathrm{sd}}$ over pairs is unchanged by resampling, at 1.000, 0.996 and 0.999 times its archived value for MMLU items, GSM8K items and MMLU subjects, while the median displacement rises by factors 1.586, 1.207 and 1.534. Resampling the numerator alone, with each pair's denominator held at its archived value, reproduces 100.3, 100.0 and 100.0 per cent of the shift in the median U; resampling the denominator alone reproduces 0.1, 0.7 and 0.4 per cent. The displacement is an absolute value, and the archived one is small relative to each pair's own item-resampling spread, with median ratio 0.856 on MMLU and 1.467 on GSM8K; adding resampling variation to such a quantity and folding it raises its median, and the rise concentrates where the archived displacement is smallest (ratios 4.47, 1.47, 1.02 and 1.00 across archived-displacement quartiles on MMLU items). Two cautions. The magnitude of the rise is not a measure of the added variation alone: the archived displacement distribution is flat near its median, so the MMLU median moves 12 percentile points and 58.5 per cent in value. And the pooled row of Table 18 admits no such decomposition, because its median displacement and its median $\widehat{\mathrm{sd}}$ are attained at different pairs; the two pooled factors multiply to 1.270 against an observed pooled ratio of 1.359. The per-design readings above are the ones that decompose.

Table 18: Conditional sensitivity of the corrected-release median U. Each bracket contains the empirical 2.5th and 97.5th percentiles of those draws. The pooled statistic takes the median after concatenating the eligible pairs. GSM8K draws are shared by the two designs.

Source	Original	Item resampling	Subject sensitivity
MMLU-Redux	0.287	[0.346, 0.598]	[0.300, 0.717]
GSM8K-Platinum	0.470	[0.425, 0.706]	[0.425, 0.706]
Pooled	0.374	[0.421, 0.604]	[0.393, 0.646]

All draws have nonempty support. The pooled draw medians are 0.508 and 0.498, above the original 0.374. These ranges describe perturbations of a nonlinear statistic within the particular model, subject and correction archives used here. Transfer to new archives remains unevaluated.

Every originally eligible pair has positive variance. In 20 GSM8K draws at least one eligible pair has zero sampled variance, always with zero displacement. Using the strict convention $U = 0$ for zero displacement and $U = +\infty$ otherwise gives the same medians as the retained denominator floor in every draw of both designs. This audit retains the separately specified control and replay conventions. Independent implementations, one using itemwise pair differences and the other binary-score cross-products, reproduce all weights, integer moments and reported summaries. They share the numerical and random-number libraries; this is computational checking of the same archived experiment.

O Re-scoring fixed modern outputs on MMLU-Pro

Question and population. The corrected keys retain the ordering of the fixed contemporary outputs. We use public Hermes and Qwen evaluation archives on MMLU-Pro (Wang et al., 2024), each containing $12{,}032$ responses. The archives postdate the early key corrections, so applying those keys measures retrospective sensitivity of the later archived outputs. We fixed both archives, four official versions, the scorer and all contrasts before reading the archived responses or scores. Earlier key-change metadata and input-matching counts were already known.

What is reproduced. We reconstruct the configured first-response metric from the Qwen archive's reported scoring-code revision. The extractor uses its original pattern priorities and case sensitivity to determine the extracted answer. Every archived per-item score is reproduced in both files. Hermes has 614 unparsed responses and Qwen has 59; these remain incorrect in the complete denominator. Full-archive accuracies are 61.902% and 70.063%, respectively. These denominators differ from the matched-key comparisons below.

The archive labels identify fixed output files. Model-weight and inference-mode identities remain unverified: the Qwen configuration and directory disagree about reasoning mode, and the Hermes scoring-code revision remains unresolved. These metadata limits persist despite numerical agreement with the common scorer.

Paired support and legal revisions. An item enters a contrast when its rendered question, ordered choices and instruction have a one-to-one match at both endpoints for both archives, with mechanically valid keys. The same physical response is used at both endpoints. Changed inputs, ambiguous matches and invalid keys remain in the accounting but are retained as excluded item classes. Table 19 reports all four overlapping contrasts on the same two model archives.

Table 19: Fixed-output MMLU-Pro comparisons. Margins are Hermes accuracy minus Qwen accuracy, in percentage points, on the same questions at both endpoints. The editable count is restricted to independently classified key-only changes within that support. Qwen remains ahead in every contrast.

Official versions	Questions	Editable	Old margin	New margin
Initial to second correction	11,712	7	-8.376	-8.359
Initial to first correction	11,712	6	-8.376	-8.359
First to second correction	12,028	1	-8.173	-8.173
Second correction to latest snapshot	11,547	0	-7.647	-7.647

From files to a decision. In the initial-to-second contrast, Hermes changes from $7{,}251$ to $7{,}256$ correct answers and Qwen from $8{,}232$ to $8{,}235$ on the same $11{,}712$ questions. The margin therefore moves by $2/11{,}712$, or 0.0171 percentage points. Its displacement is 0.039 times the original empirical item-sampling standard deviation. This is a descriptive scale for the fixed pair and questions used here.

Allowing every key-only item in that support either to take any legal answer or to be deleted jointly for both models gives an exact margin range of $[-8.387, -8.333]$ percentage points. We obtain it by enumerating the retained editable count and summing the smallest or largest attainable item contributions, as in Appendix N. The whole range is negative, so the comparison is certified to retain Qwen under this declared class. The observed later keys belong to the class. The edit set is identified from the observed releases, so this is a conditional certificate given that set.

The latest-version contrast has zero valid key-only changes on its paired support. The complete official dataset also contains input edits and a formerly invalid key outside this fixed-input comparison. These results describe contemporary archived outputs under the four specified key contrasts.

Version identifiers. The official dataset commits, in table order of time, begin `014c8cf52021`, `295c2820d698`, `475d58ba0cc1` and `b189ec765aa7`. The Hermes and Qwen archive commits begin `904a45855244` and `42f1596e0dbf`; the located scoring-code commit begins `59fe31802159`. Full byte hashes and per-item tables accompany the computation records. The derived tables contain scores and identifiers, with raw response text excluded.

P WHOLE-PANEL WINNERS UNDER REVISED KEYS

The original winner survives every observed common-support revision. For each common-support event, we score every available model under both keys and compare the complete maximizer sets. The eight events in Table 20 each have a unique original winner that remains the unique optimum. The survey series use pair-specific observed quarters and therefore support pairwise comparisons on different observation sets.

Table 20: Whole-panel decision stability for the eight common-support events. All original winners remain uniquely optimal under the revised key. The events overlap in models and data.

Event	Models	Original items	Revised items	Winner changed
MMLU to Redux	59	5,381	5,127	0
GSM8K to Platinum	60	1,319	1,209	0
SWE-bench to Verified	24	2,294	500	0
SWE-bench to Lite	24	2,294	300	0
TIME Job Claims	26	104	104	0
TIME US Labor	26	240	240	0
TIME Housing Inventory	26	144	144	0
TIME Vehicle Supply	26	300	300	0

For MMLU, the global original-key winner Yi-1.5-34B-32K remains optimal under both exact prompt identity and whitespace-normalized identity. On the primary retained support it has $4{,}067$ original-key and $4{,}107$ revised-key correct answers out of $5{,}127$. Subject-level choice can still change: the full-support contrast has positive original-winner regret in six subjects, versus two when both keys use the retained support (Appendix N). These are observed score comparisons under the released keys; stability under an unknown future key requires a stated revision budget.

Q A SIZE-MATCHED REFERENCE FOR CURATED TASK SETS

Fixed outputs and support. We use the same archived submissions as the main SWE-bench analysis. For each submission, an official task receives score one if its identifier occurs in the archived resolved list and zero otherwise. Repeated identifiers contribute the same binary membership score. One source lists seven resolved identifiers repeatedly, with twenty-eight extra entries; these are retained in the input diagnostics. Every resolved identifier belongs to the full official support. Failed and ungenerated tasks remain in the denominator.

The full set has $N = 2{,}294$ tasks. Verified retains $m = 500$ and Lite retains $m = 300$. All 24 submissions and their 276 unordered pairs are kept for the subset-selection tests. One original tied pair per subset is excluded from the 275-pair strict-order summaries used in the figure and abstract. The all-pair medians are 0.5763 and 2.7404; the strict-order medians are 0.5813 and 2.7462. This is a denominator distinction within one comparison set with unchanged models and scoring policy.

Discrete reference distribution. For pairwise contributions $v_i \in \{-1, 0, 1\}$, write n_+ and n_- for positive and negative counts. Uniformly choosing m tasks gives selected counts $(X, Y, m-X-Y)$ with law

$$\Pr(X = x, Y = y) = \frac{\binom{n_+}{x}\binom{n_-}{y}\binom{N-n_+-n_-}{m-x-y}}{\binom{N}{m}}. \tag{17}$$

The subset margin is $(X - Y)/m$. With $s = n_+ - n_-$ and observed subset sum s_m, the two-sided tail includes every state satisfying $|N(X-Y) - ms| \geq |Ns_m - ms|$. Integer comparisons preserve equality at the observed boundary. The law is exact for uniform sampling without replacement; probabilities are evaluated numerically from the discrete law. Equation 6 gives its variance; tail probabilities additionally require the probability law.

We use one Holm family for all pairs and both subsets. This controls familywise error under the stated conditional null hypotheses despite dependence between pairs. The curators' selection mechanism and the causal share of displacement are outside this measurement. A retained null can be compatible with both sampling variation and systematic selection.

Crossed resampling sensitivity. Pairs share models and tasks. Following the row-and-column resampling principle of Owen (2007), each replicate samples model identities with replacement and samples tasks within the four joint Verified/Lite membership strata. The same task multiplicities apply to every model and both subsets; stratum sizes remain fixed. A pair receives the product of its two model multiplicities. Self-pairs are excluded, and the original nonzero-margin mask is held fixed when reporting strict-order summaries. Every margin, disagreement count and sampling

variance is recomputed. The displayed interval uses percentiles of 2,000 crossed model-and-task replicates.

These intervals describe sensitivity under the empirical archive and fixed membership strata. The submissions are the available archive, with shared training lineages and dependence between issues from the same software repository. The resampling design conditions on the archive without modelling these additional dependencies. Population coverage for the median and transfer to future models remain unestablished; crossed-mean theory concerns a different statistic.

Implementation checks. Separate implementations reconstruct the raw matrix and compute the complete tables. One evaluates a conditional hypergeometric tail and the other sums joint category probabilities. The matrices, counts, decisions and all resampling summaries agree; numerical probability checks use a predeclared absolute tolerance. A failed initial parser assertion concerning repeated success identifiers is retained. Its correction leaves binary membership scores, the model set, the task set and the statistical specification unchanged.

R REGISTERED TEST OF EDIT PLACEMENT

The protocol was frozen after the displacement results were known but before its placement statistics were computed and is distributed with the supplement. It fixes three panels, the per-item statistics, the null, three hypotheses, the multiplicity rule and the decision rule.

Panels and groups. Panel A is the corrected multiple-choice release under exact-prompt identity (59 models, 5,381 items, 254 removals, 99 key corrections); panel B is the same release under whitespace-folded identity (5,651 items, 263 removals, 100 corrections); panel C is the corrected grade-school arithmetic release (60 models, 1,319 items, 110 removals, 10 corrections). Corrected items are those retained with a changed gold answer; removed items are those absent from the analysed revised support. On panels A and B these are our annotation-status exclusions, not deletion instructions from MMLU-Redux; panel C uses the published arithmetic support. Appendix N specifies the MMLU policy.

Statistics and null. For item i, c_i is the fraction of models scored correct under the original key, with invalid or missing answers scored incorrect; $v_i = c_i(1 - c_i)$ and $w_i = \mathbf{1}\{c_i \leq 0.2\}$. Neither uses the revised key. For a group of size k the null draws 10,000 uniform subsets of size k from the panel and records the median of v and the mean of w. Because $c_i = j_i/M$ takes finitely many values, v_i is carried as the integer $j_i(M - j_i)$ over M^2 and a median as an integer over $2M^2$, so every comparison with the null is exact; evaluating $c_i(1 - c_i)$ in floating point makes j and $M - j$ differ in the last place and reorders tied draws.

Outcomes. Reported p-values are one-sided Monte-Carlo values $(1 + \#\{\text{null at least as extreme}\})/(B + 1)$, Holm-adjusted across the three panels within each family.

H-A and H-C hold on all three panels. H-B holds on none: removals sit above the null on the two multiple-choice panels without reaching the registered level, and below it on the arithmetic panel. The registered family is the three panels within each hypothesis, so the table adjusts over three. Adjusting instead over all nine tests at once changes no decision: the six rejected values become 0.0009, 0.0009, 0.0009, 0.0009, 0.0010 and 0.0028, and the three H-B values are unchanged at 0.2514, 0.2514 and 1. Six of nine reject under either family. The protocol's decision rule required both H-A and H-B on at least two panels before a combined statement about corrections and removals could be made, so this paper reports the placement of corrections and reports the removal test as executed.

Identity checks. For all $1{,}711 + 1{,}711 + 1{,}770$ model pairs, the archived margin sums are reproduced from the score matrices in integer arithmetic with no discrepancy; the relabelling identity reproduces every revised margin from the corrected items alone, with every non-corrected contribution identically zero; and the removal identity reproduces every reduced-support margin. Wherever two models emitted the same parsed answer, both their score difference and its change are exactly zero, on all pairs.

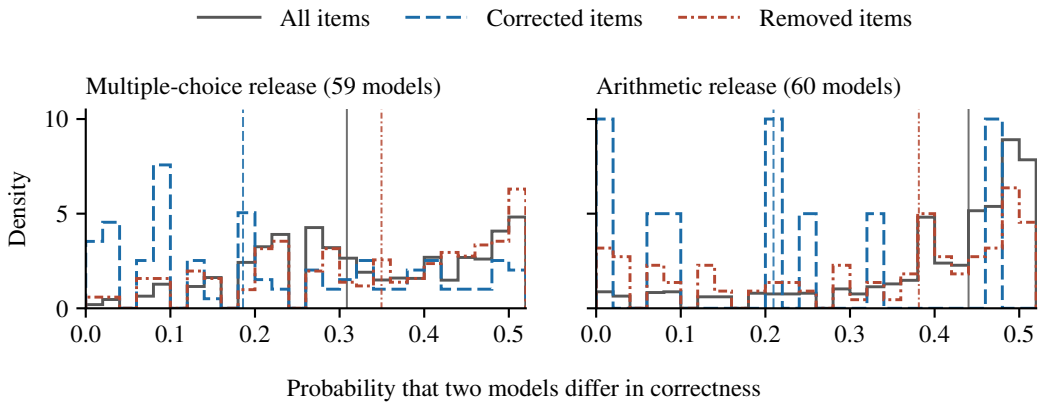


Figure 3: Where a revision lands. The horizontal axis is $2c(1-c)M/(M-1)$, the probability that two distinct uniformly drawn models differ *in correctness* under the original key; c is the fraction correct among M models. Densities include all items, including probabilities above 0.5 in a finite roster. Vertical lines mark group medians. MMLU removals are our annotation-status exclusions, not maintainer deletions. Corrected items concentrate to the left of the panel as a whole; removed items do not separate in a common direction. Answer agreement is a different event.

Table 21: Registered placement tests. H-A asks whether corrections sit below the null on model disagreement, H-B whether removals sit above it, H-C whether corrections concentrate on items most models answer incorrectly. Null medians are lattice values: with $M = 59$ models, v takes the values $j(M-j)/M^2$, and the median of a size-k subset is the atom $11 \cdot 48/59^2 = 0.1517$ for odd k and the midpoint of atoms 11 and 12, 0.1569, for even k; the two MMLU panels therefore show the same two values with the roles of k reversed, not a row swap.

Family	Panel	k	Observed	Null median	Raw p	Holm p
H-A (median v, corrections)	A	99	0.0914	0.1517	0.0001	0.0003
	B	100	0.0914	0.1569	0.0001	0.0003
	C	10	0.1031	0.2193	0.0007	0.0007
H-B (median v, removals)	A	254	0.1718	0.1569	0.0937	0.2514
	B	263	0.1718	0.1517	0.0838	0.2514
	C	110	0.1875	0.2193	1.0000	1.0000
H-C (mean w, corrections)	A	99	0.6465	0.1313	0.0001	0.0003
	B	100	0.6500	0.1300	0.0001	0.0003
	C	10	0.8000	0.2000	0.0002	0.0003

Sensitivity and independent execution. Drawing the correction null from retained items only leaves every H-A and H-C conclusion unchanged. A separate implementation, written from the frozen protocol without sight of the first, reproduces every observed statistic and every decision in this appendix, and repeating the null with forty independent master seeds leaves the decisions unchanged.

The same scorer contrast on the multiple-choice panel. Holding the archived answers, the item set and the original key fixed, we replaced the panel's own archived correctness flags with the declared parser used throughout this paper. On the 59 models and 5,381 items of the multiple-choice panel the two scorings agree on every model–item cell: no score changes and none of the 1,711 pairwise comparisons reverses. The margins under the declared parser reproduce the panel's archived margin sums exactly in integer arithmetic. The corresponding substitution on the free-text arithmetic panel changes 2,699 cells and reverses 123 comparisons. This contrast identifies a difference between archived scoring pipelines, not a general effect of free-text versus multiple-choice answers. The protocol for this contrast was registered before it was computed and travels with the supplement.

Agreement in correctness against agreement in answer. The registered statistic uses agreement in correctness under the original key, which is observable before the revision is known. The relabelling identity, however, vanishes on agreement in *answer*. These differ, and on this collection they point in opposite directions: corrected items have high agreement in correctness and slightly lower agreement in answer than the panel as a whole (0.538, 0.541 and 0.175 against 0.561, 0.562 and 0.254). Counting the identity directly, a single correction changes the margin of 0.427 and 0.424 of all model pairs on the two multiple-choice panels, with 28% of models answering neither the old nor the new key. The placement result is therefore reported as a property of where corrections are made, and the displacement comparison rests on the matched control rather than on this statistic. This companion calculation was made after the registered test and is descriptive.

R.1 Selection-aware legal placement controls

This exploratory analysis was specified after the earlier displacement and uniform-control results were known. It uses the exact-prompt MMLU panel on its $5{,}127$ retained items, holding every prediction, original margin, sampling spread and the 99 genuine label changes fixed. There are $1{,}711$ model pairs, of which $1{,}708$ have a nonzero original margin on this support. These denominators differ from the full-support comparisons in the main table.

Within each stratum we relocate the observed number of edits and permute its observed new-label tokens. Subject and original label are fixed, so every assigned label is legal, differs from the original, and is shared by all models. Both specifications also match the number of missing predictions. The fine specification matches the exact original correct-model count and number of valid-model pairs giving different answers. The coarse specification bins the corresponding fractions into fifths, with endpoints fixed before execution. Neither specification uses future displacement to select bins. Each draws 200 assignments from its declared legal set.

Exact matching leaves only four legal keys and fixes 97 of the 99 edit locations. All draws have six strict reversals. Coarse matching permits 51 edits to move and relocates a median of 31 per draw. The observed median displacement is 0.2197; the coarse reference has median 0.2384 and 5th–95th percentiles $[0.1936, 0.2903]$. Its median reversal count is six, with percentiles $[4, 9]$, compared with six observed reversals. No strata or degenerate edits are dropped. The new reference substantially narrows the earlier unstratified contrast. Unknown curator selection probabilities and limited overlap prevent a causal interpretation. Model pairs share items and models; the randomization unit is one complete legal key, not a model pair.

R.2 Key, scoring contract and support

We rescore the same $60 \times 1{,}319$ frozen arithmetic outputs under four explicit contracts, two keys and two supports. The revised support contains $1{,}209$ items. No model is regenerated. For the 110 removed items, the full-support new-key arm carries forward the original answer. This defined counterfactual supports the decomposition; it is not the published revised benchmark. All cells are binary under the specified scoring programs; rejected parses remain zero rather than being removed. The public strict rule extracts the first number following the prescribed answer marker; the flexible rule extracts the last number whether or not that marker appears. Their pinned public implementation applies the same string-normalization metric. The historical strict rule extracts the first marked number with comma removal and exact string comparison. Our common parser takes the last marked answer and attempts numeric normalization before string comparison. The supplement records the source commits, functions, synthetic examples and all factorial contrasts.

The historical strict rule reproduces every archived original-key flag. The current public strict rule differs on one cell; the flexible rule differs on $3{,}245$, and the common parser on $2{,}699$. The historical rule's behavior is therefore sufficient to reconstruct these flags, but the archive does not identify its actual historical dependency revision.

At the original key and support, public strict versus flexible scoring changes $3{,}244$ cells and reverses 153 pairs. Historical strict versus common numeric scoring changes $2{,}699$ cells and reverses 123 pairs. An ordered structural classification assigns these cell differences to missing or rejected marker format ($2{,}008$), first versus last marker (325), numeric-string normalization (155), suffix or expression handling (69), and other format differences (142). These categories partition cells, not

Table 22: Strict pair reversals with one scoring contract fixed. Key-only keeps the full original support; removal-only keeps the original key; combined changes both. All 1,770 unordered pairs are evaluated and original ties do not count as reversals. Counts need not add: two changes can reinforce or cancel. The main tables keep the common parser rather than select a scorer after observing these results.

Scoring contract	Original ties	Key only	Removal only	Combined
Historical strict	16	2	9	10
Public strict	16	2	9	10
Public flexible	4	1	6	9
Common numeric	8	1	7	8

pair reversals, and do not establish semantic correctness. For example, comparing the synthetic strings `10.0` and `10` exposes numeric versus string equality without deciding whether either occurs as a valid solution in an actual response. The archived main estimand uses the same common parser under both keys and remains unchanged.

An exploratory coherence check holds the original key and all 1,319 items fixed. For each pair oriented toward the candidate historical rule's leader, a whole-rule test requires a positive score gap under each of the four policies; a conservative relaxation also permits the policy to vary independently by item. Of all 1,770 pairs, 16 tie under the candidate historical rule and remain unclassified. Among the other 1,754, the whole-rule test certifies 1,577, against 1,512 for the relaxation: 65 pairs pass only when a single complete rule applies throughout. The unique candidate-historical leader passes both tests against every challenger. The rule was fixed before this derived count, but the source and policies were already examined. Finite policy enumeration is classical; these counts establish neither a sequential verification gain nor a change in model-selection loss.

The scope of this constraint matters. The pinned public GSM8K task[1] applies each named filter pipeline to all its responses and reports strict-match and flexible-extract metrics separately. These are two reported metrics, not a maintainer-declared choice among all four rules above. The historical candidate matches the archived flags without identifying their runtime; the local parser is an analysis specification. Other tasks may use different filters, and a new document-dependent filter would define another policy. Thus the four complete rules form an author-declared sensitivity set. The itemwise relaxation lies outside that set, not outside every evaluation procedure the software permits.

Table 23: Three witnesses to the itemwise relaxation gap, selected by system-index order before inspecting their gaps. Entries are sums of oriented score differences over the same 1,319 items, not percentages. Every complete rule gives a positive sum, whereas selecting an adverse rule separately for each item gives a negative sum. Indices refer to the released derived score table.

System pair	Public strict	Public flexible	Historical	Local	Itemwise
$(3, 4)$	5	5	5	1	-7
$(3, 24)$	8	7	8	8	-7
$(3, 52)$	85	67	85	3	-15

For each witness the released item-level assignment attains the relaxed sum. It does not coincide with any of the four complete rules. Adding a permitted item-type router would enlarge the sensitivity set and require another check; the present count makes no claim over such an enlargement.

An independent implementation rereads all raw outputs, reproduces every cell under all four contracts and both keys, and matches every pair contrast. The analysis nevertheless concerns one source and follows exposure to its earlier scorer discrepancy. Structural audits are not human gold annotations, and behavioral reconstruction is not proof of the original software environment.

[1]The task configuration declares both named filter pipelines.

R.3 DO DIFFERENT SCORING RULES REVERSE THE SAME COMPARISONS?

The four rules produce similar numbers of reversals under the combined answer and support update, but counts alone do not identify the affected comparisons. We extend the existing factorial with a post-exposure descriptive census. Every calculation uses the original integer model ordering, $a < b$, over all $\binom{60}{2} = 1{,}770$ pairs. A strict reversal has opposite nonzero endpoint signs; a zero margin is a tie and remains in the pair table. For each rule r, let F_r be its reversal set for the original key on $1{,}319$ items versus the revised key on the retained $1{,}209$ items.

These sets have sizes 10, 9, 10 and 8, a union of 22 and an empty intersection (Table 24). Their intersections with the separate 153-pair strict-to-flexible scorer contrast have sizes 5, 2, 5 and 2 respectively. Thus the eight-to-ten counts are not subsets of those 153 pairs. Both analyses use the same comparison population, but change different evaluation components. Across all rules, 25 pairs have at least one tied endpoint; they are retained rather than silently dropped. The complete supplement records every endpoint sign and set membership.

Table 24: Overlap of combined answer/support reversal sets. S and F are public strict and flexible scoring; H is historical strict and N is the common numeric parser. Jaccard overlap is intersection divided by union. All comparisons refer to the same $1{,}770$ pairs.

Rule pair	S–F	S–H	S–N	F–H	F–N	H–N
Intersection	1	10	3	1	1	3
Union	18	10	15	18	16	15
Jaccard	$1/18$	1	$1/5$	$1/18$	$1/16$	$1/5$

Table 25 separates changes to answers and support. Full-support new-key calculations carry the old key forward on the 110 removed items; this is a decomposition convention, not a released evaluation arm. The original main-table denominator of $1{,}762$ non-tied arithmetic pairs removes the common parser's eight original ties, whereas this census retains all $1{,}770$. Neither denominator counts independent maintenance events.

Table 25: Strict reversals under each fixed rule, and overlap across rules. Key-only fixes the stated support; support-only fixes the stated key. Zero endpoint margins never count as reversals. Counts may cancel or reinforce when operations are combined.

Operation	S	F	H	N	Union	Intersection
Key and support	10	9	10	8	22	0
Key only, full support	2	1	2	1	4	0
Key only, retained support	1	3	1	1	5	0
Support only, old key	9	6	9	7	17	1
Support only, new key	8	8	10	7	21	0

For a fixed support T, write $d_{ab}(k, r; T)$ for the difference of the two models' mean binary scores. The classical key-by-scorer interaction is

$$I_{ab}^{r_0,r_1}(T) = \big[d_{ab}(k_1, r_1; T) - d_{ab}(k_0, r_1; T)\big] - \big[d_{ab}(k_1, r_0; T) - d_{ab}(k_0, r_0; T)\big].$$

We compute an integer numerator over $|T|$, without rounding margins before comparing signs. Table 26 reports every scorer pair. Since removed items carry the old key, the interaction numerators are identical on both supports; only the denominator changes. All median absolute interactions are zero. Small interactions can still matter for close pairs, and the initial scorer-specific margins also affect whether a key update reverses a comparison. In particular, zero interaction does not require the initial margins to be equal.

As concrete examples, the first pair in numeric index order with reversal pattern S/H/N is $(0, 12)$: its strict numerator changes from -2 to 4, whereas the flexible numerator changes from 1 to 6. For the first F-only pair, $(3, 54)$, the flexible numerator changes from 7 to -7, while the strict numerator changes from 24 to 9. The first N-only pair, $(18, 52)$, changes from -2 to 2 under N and from -88 to -78 under S. Old and new denominators are $1{,}319$ and $1{,}209$. These examples follow a declared index-order selection; the model registry and all pairs are supplied, including ties.

Table 26: Fixed-support key-by-scorer interactions over all 1,770 pairs. Absolute 95th percentiles use the nearest-rank definition. Multiply the displayed numerators by $1/1{,}319$ on full support or $1/1{,}209$ on retained support. The last columns count pairs for which the two scorers disagree on whether the key-only update reverses order.

Rules	Nonzero	$\lvert T\rvert\, q_{.95}(\lvert I\rvert)$	$\lvert T\rvert \max \lvert I\rvert$	Full	Retained
S–F	386	1	3	3	4
S–H	0	0	0	0	0
S–N	174	1	3	3	2
F–H	386	1	3	3	4
F–N	388	2	4	2	4
H–N	174	1	3	3	2

The calculation protocol and implementation were fixed before computing these joint summaries, but the archive and earlier marginal counts were already examined. A separate implementation rebuilt the support mask from the original sources and matched every pair, overlap and exact interaction. The results describe scorer-dependent revision judgments in one archive; they are not an equal-scale causal contrast, new certification method, semantic validation or an independent event replication.

S Displacement in raw units

Which sources account for the reversals. Of the original collection's 458 reversals, 437 fall in routine survey and statistical revisions, 427 in the four survey horizons and 335 in the two real-GDP horizons. The two real-GDP reversal rates are 20.4% and 20.6%. The four benchmark series contribute 21 reversals, at rates of 0.41%, 0.45%, 0.73% and 1.45%. These source-specific rates answer different questions from a pooled rate. Of all 458 reversals, 391 start within one item-sampling standard deviation; an independent-item Gaussian reference gives approximately 694. That reference is a descriptive yardstick, not a model of how maintainers choose corrections. Twenty-two observed reversals cross the one-sided sampling-reference threshold, twenty-one in real-GDP comparisons.

The main text reports displacement in units of each comparison's own item-sampling spread, which is what makes revisions of different sizes comparable. Table 27 gives the median raw displacement $|d'-d|$ beside it for every series. The two multiple-choice and arithmetic releases and the two curated subsets are all measured as fractions of the scored items and are comparable with one another; the survey and economic series are differences of forecast losses in their own units and are not comparable with either group.

Table 27: Median raw displacement and median displacement in item-sampling units, with the number of scored items. Raw values are comparable only within a scale. For the survey series the count is the median number of quarters a pair is scored on, not a common support: the calendar covers 140 quarters and individual pairs are scored on 20 to 95 (Appendix E.1).

Series	Scale	Items	Median $\lvert d'-d\rvert$	Median U
MMLU-Redux	fraction of items	5,381	0.0018	0.29
GSM8K-Platinum	fraction of items	1,319	0.0067	0.47
SWE-bench Verified	fraction of items	2,294	0.1010	12.47
SWE-bench Lite	fraction of items	2,294	0.0587	7.08
SPF GDP, $h=0$	forecast loss	42	0.0873	0.65
SPF GDP, $h=1$	forecast loss	42	0.0674	0.48
SPF unemployment, $h=0$	forecast loss	43	0.0064	0.36
SPF unemployment, $h=1$	forecast loss	42	0.0048	0.16
TIME Job Claims	forecast loss	104	0.0286	0.19
TIME US Labor	forecast loss	240	0.0035	0.08
TIME Housing	forecast loss	144	0.0000	0.00
TIME Vehicles	forecast loss	300	0.0006	0.02

Pooled by operation, the median raw displacement is 0.0040 for corrected releases and 0.0781 for curation on the shared item-fraction scale, and 0.0091 for routine statistical revisions in loss units. The ordering of the two operations that share a scale is the same in raw units as in sampling units. The sampling unit divides by a spread that falls with the number of scored items and rises with how often the two models differ, so the two columns are not proportional.

T SEPARATING ANSWER AND DATABASE REVISIONS IN TEXT-TO-SQL

Fixed outputs, distinct interventions. The text-to-SQL correction studies of Jin et al. (2026a;b) change questions, reference queries and databases, and rerun agents on the corrected inputs. We instead retain their released original SQL: sixteen systems on the same hundred question identifiers, giving $1{,}600$ fixed outputs. Each is evaluated under both reference queries on both database versions. The four evaluations separate reference-query replacement on the original database, database replacement with the original reference, and their joint replacement. No new SQL is generated, and no failed output is removed. This is a retrospective analysis of one curator's release, whose published aggregate results were already known; the two papers are not independent sources.

We join by question identifier rather than by row position. Before execution, we fix the outputs, reference queries, database hashes, resource limits and question-level strata. Reference-query text changes on 41 questions; questions or evidence change on 44; four of eleven databases change bytes. There are 56 questions with unchanged question and evidence text, 63 with unchanged database bytes, and 37 satisfying both. Every corrected schema directory adds a file absent from the original directory. Available schema files are therefore not identical, and the archives do not establish whether each agent used that additional file. The strata cannot establish full prompt identity or the semantic correctness of a revised reference query.

Execution contract and unresolved outcomes. The primary contract compares sets of returned rows, following the released evaluator. Multiset equality and exact row-order equality are separate sensitivities, not claims that preserving order is always appropriate. Queries run in isolated, read-only SQLite processes, with a ten-second wall limit, 512 MiB memory, $10{,}000$ returned rows and 4 MiB serialized output. Unsupported or time-dependent operations, reference-query failures, resource limits and prediction SQL errors remain unknown. In particular, a failed execution is not silently scored as an incorrect answer. All denominators remain fixed. Of $6{,}400$ evaluations, $5{,}884$ yield scores and 516 remain unknown. A separate implementation executes the archived databases and matches all $19{,}200$ scores across the three equality contracts, including unknown status. It does not provide human semantic adjudication.

Table 28: Why the fixed SQL replay remains partially scored. Counts cover all four reference/database cells. A blocked reference prevents scoring the corresponding prediction. Additional time under the unchanged contract resolves none of these causes.

Recorded cause	Reference	Prediction	Required change
Returned-row cap	256	8	New resource contract
Clock-dependent operation	192	0	Defined temporal target
SQLite query error	0	40	Query/schema diagnosis
Function/authorizer rejection	0	16	Qualified scorer change
SELECT/WITH filter rejection	0	4	Output/parsing diagnosis
Total	448	68	

These 516 cells share 38 blocking execution objects; none failed because of a timeout, exhausted source budget or recorded infrastructure error. The frozen runner caches failures as well as successes and does not retry them. No known score with the same complete prediction/reference/database identity fills an unknown cell. A changed row cap or clock convention would be a new endpoint, with unmeasured cost. The archived SQLite reasons omit the original exception text, so query-error subtypes cannot be recovered from the logs alone. An output rejected by the SELECT/WITH filter is not necessarily a mutating query. In all 448 reference failures, the runner did not request the prediction; the other 68 unknown scores have failed predictions. No unknown score has both

complete result objects, so an additional shared-reference equality cannot score these cells under the original execution contract.

Identification without discarding failures. For a model with c correct and u unknown scores on n fixed questions, accuracy lies in $[c/n, (c+u)/n]$. Thus the difference between models a and b lies in

$$\left[\frac{c_a - c_b - u_b}{n}, \frac{c_a + u_a - c_b}{n}\right].$$

A strict reversal is established when the entire old interval is positive and the entire revised interval is negative, or conversely. These are deterministic bounds over missing binary outcomes, not sampling confidence intervals. They describe all binary extensions of the partial score table; an extension does not repair an undefined reference or make a failed query executable under the original contract. Ignoring dependencies among unknown scores can widen the bounds but cannot create a false established reversal. The supplementary tables retain every model pair; they do not rank models by accuracy conditional on successful execution.

Table 29: Fixed-output SQL replay under set equality. Gains and losses are known changes from incorrect to correct and conversely. Unknown denotes an unresolved transition, not a zero score. Established reversals hold for every binary completion of the unknown scores; they are lower bounds, not exact reversal totals. Each row contains all 120 model pairs. The last two rows are predefined input strata, not additional independent releases.

Intervention	Outputs	Gains	Losses	Unknown	Reversals
Reference only	1,600	189	153	129	≥ 6
Database only	1,600	2	6	129	≥ 0
Reference and database	1,600	189	170	129	≥ 5
Reference, unchanged question/evidence	896	98	24	44	≥ 15
Reference, unchanged text and DB	592	49	12	22	≥ 3

On the full support, replacing the reference alone changes 342 known scores; replacing only the database changes eight. Even with unresolved executions, at least six pairwise orderings reverse under reference replacement. On the 37 questions with unchanged question, evidence and database bytes, 61 known scores change and at least three orderings reverse. The full-support result includes corrected references for changed questions; it measures reference replacement, not necessarily correction of an answer to an identical question. The restricted result still lacks a complete prompt trace. Neither establishes a new model's superiority or generalization to another correction policy.

Displacement without conditioning on execution success. To use Equation 1 with unknown scores, let $S = \sum_i g_i$, $Q = \sum_i g_i^2$ and $S' = \sum_i g'_i$. Then

$$U = \frac{|S' - S|}{\sqrt{Q - S^2/n}}, \qquad \frac{|d|}{s} = \frac{|S|}{\sqrt{Q - S^2/n}}.$$

For each item, unknown binary scores permit a finite set of contributions $g_i \in \{-1, 0, 1\}$. A dynamic program enumerates attainable (S, Q); the independent revised contributions give an integer interval for S'. Minimizing and maximizing over their product gives conservative bounds. This relaxation ignores shared-query constraints, so its endpoints need not be jointly achievable across evaluation versions or model pairs. In particular, identical databases need not change any unknown score, whereas the relaxation allows them to. These are deterministic outer bounds, not confidence intervals or a new identification theorem. No pair in this calculation has possible zero original spread; the program nevertheless records explicit conventions for that case. It covers all 4,320 pair–stratum–contract–intervention combinations, without changing the frozen scores.

For Contextual-SQL minus MAC-SQL on all 100 questions, original margin sums range from 6 to 21 and revised sums from -16 to -1. Under set equality, the original $|d|/s$ is at least 1.2340, and U lies within the outward-rounded interval $[1.43, 8.16]$. This establishes a reversal from beyond one sampling unit even under the broader independent-completion relaxation. With unchanged question and evidence text, the same pair has old sums $[4, 9]$, new sums $[-8, -3]$ and $U \in [2.52, 6.19]$.

The unknown consumed prompts still prevent interpreting this as a controlled same-task semantic correction.

Taking medians of the pairwise lower and upper bounds gives a conservative envelope for median displacement. On all questions under set equality, these envelopes are $[0, 5.24]$ for reference replacement, $[0, 3.71]$ for database replacement and $[0, 5.22]$ for their combination, rounded outwards. Their width does not support a precise comparison of the SQL median with the original collection's correction median. The identified reversal above is the stronger conclusion. Neither success-only rankings nor imputed zero scores are used.

Sharing one archived execution object. A further sensitivity retains equality between repeated references to the same cached score target. Identity includes the question-local cache, prediction SQL, reference SQL, database bytes, scoring code, resource limits and set-equality contract. We exclude clock-dependent or unsupported operations and cells lacking an archived execution object. Across the four evaluation cells this admits 52 unknown scores, of which 44 share a class with another score. The reference-only comparison below contains 258 unknown scores, with 26 admitted and 14 in shared classes. All 516 outcomes remain unobserved: assigning a common binary value to a class is a completion assumption, not a successful query replay.

For each item and model pair, enumerating the permitted score bits gives tuples (g_i, g_i^2, g_i'). A finite-state recurrence sums these tuples over items, preserving old–new dependence. Extrema of U and $|d|/s$ over the resulting (S, Q, S') states are exact within this completion abstraction. They need not be attainable as successful SQL outputs. This is a classical enumeration, not a new identification theorem. Exhaustive small-instance checks precede the archive calculation; an independent implementation checks the identities and extrema.

On all 120 pairs under reference replacement and set equality, the shared constraints tighten 8 displacement lower bounds and 65 upper bounds. The original and revised margin intervals, their identified strict signs, original-margin standardized bounds and established reversal decisions do not change. Neither abstraction identifies a strict champion; both give the same worst-case regret for each model. For Contextual-SQL minus MAC-SQL, the displacement envelope narrows from $[1.43, 8.16]$ to $[1.61, 8.05]$, rounded outwards; the original standardized-margin lower bound remains 1.2340. The narrower displacement interval therefore supplies no additional model-selection decision here.

A full-roster projection gives a stronger negative result: of the 129 unknown revised scores, 13 enter eligible shared classes, but none is tied to another revised score or fixed by an observed score. Consequently the shared and independent binary completions have exactly the same set of revised score vectors across all sixteen models. The tightened displacement bounds use dependence between old and revised scores; they cannot improve a champion, top-three or regret decision based only on revised scores.

The archived contract supplies no new query action that resolves these terminal failures. We therefore stop before an adaptive-query comparison. Any future comparison must declare a new legal action and its failure behavior, match the information and risk guarantees, and charge both methods for shared setup and cached execution. The present result is a boundary of what existing evidence identifies, not a measured reduction in verification cost.

The main execution uses $2{,}119$ subprocess queries for $12{,}352$ requests, reusing byte-identical SQL/database evaluations within each question. Its recorded query wall time is 160.87 seconds and batch time is 167.17 seconds. These are execution measurements, not method savings: download, input preparation, verification and development are additional costs. There is no pre-revision edit budget for this event, so we do not estimate one from the realized changes, run a budget-selection contest, or count nominal reuse as a prospective guarantee. The public supplement reproduces the derived tables without redistributing the databases or original generated SQL.

U AI-ASSISTED INSPECTION OF SCORER DISAGREEMENTS

Selection and masking. We inspect all 96 previously selected arithmetic responses, covering 44 models and 89 distinct question–response combinations. Selection takes one case per available model and discrepancy type between the historical strict rule and our declared parser. This delib-

erately enriched sample is neither random nor a held-out contest for choosing a scorer. We register the audit before annotation, without changing the cases or any recorded scorer output. Two fresh-context AI sessions independently read the original questions and responses, masked to model identities, scorer outputs and each other's labels. Each freezes its judgments before opening randomized reference proposals. Both sessions use the same assistant family; the exact serving revision is not exposed. This is not cross-model replication or human annotation.

Separate judgments, explicit uncertainty. Question specification, response fidelity, final-answer correctness and reasoning validity are distinct labels. Initial answer judgments agree on 94 cases. A third AI session inspects all 24 cases flagged by disagreement, ambiguity or a possible error; the coordinating session reviews its decisions against the original material and the registered rules. Original labels remain unchanged. This adjudication is not a further blind replication. A binary proxy is assigned only to well-posed questions with determinate answers: 80 are judged correct, six incorrect and ten remain undetermined. Eight correct final answers contain invalid intermediate reasoning. A response can therefore pass an answer-only check without providing a valid derivation.

Underspecified questions retain uncertainty even when a response supplies a valid conditional formula. For example, knowing how many flowers were ordered does not determine an outstanding delivery without knowing how many have already arrived. Conflicting final claims also remain unresolved rather than selecting whichever number matches a reference. These rules produce eight undetermined cases from question ambiguity or missing information and two from conflicting responses to well-posed questions. Among old references, the audit labels 88 valid, one invalid and seven undetermined. Among revised references, 78 are labeled valid, six undetermined and twelve unavailable; absence is not an incorrect answer.

Table 30: Frozen old-key scorers compared with AI-assisted judgments on selected disagreements. Accept and reject denote the scorer's binary decision, not human-validated correctness. Every row retains the full denominator. The second panel collapses identical question–response combinations; neither panel consists of independent revision events or estimates population accuracy.

Scorer	Agreement	Accept, AI incorrect	Reject, AI correct	Undetermined
All 96 selected model–case records				
Public strict	31	6	49	10
Public flexible	52	0	34	10
Historical strict candidate	31	6	49	10
Declared parser	55	0	31	10
Distinct question–response combinations (89)				
Public strict	29	5	46	9
Public flexible	48	0	32	9
Historical strict candidate	29	5	46	9
Declared parser	51	0	29	9

What the audit does not establish. Table 30 compares all frozen scorers without selecting a winner or changing the main analysis. Agreement between AI sessions does not validate their shared assumptions, and prior model exposure to the benchmark cannot be excluded. The judgments are inspectable diagnostic evidence, not semantic gold. The supplement releases derived labels, input hashes and a counting program, not original generated text. An independent implementation checks the joins and counts; this verifies computation, not the truth of the AI judgments.

V A CHECKER REVISION ON FIXED PROGRAM OUTPUTS

LiveCodeBench evaluates generated programs against executable tests (Jain et al., 2025). A public January 2025 change to its code-generation checker altered how a program's call to `exit` was handled (LiveCodeBench developers, 2025). We isolate that checker edit, not the other runner and diagnostic changes in the same commit. Four archived scoring modules match the repository state before the model-output archive byte for byte. The historical dependency image is unavailable, so matching archived scores is an additional check rather than proof of the complete historical environment.

Two archived model artifacts each provide ten frozen completions on the same 713 questions from the pinned Lite release. Before scoring this secondary endpoint, we specified the first completion for every model–question cell, matched all question identifiers and four test files, and held each generated program and its tests fixed between checkers. The original ten-completion replay is a separate endpoint. This registration followed a post-exposure single-question mechanism check. An input-digest typo blocked the first secondary launch before either checker loaded; a recorded prescore correction changed only the digest guard. No item or completion was selected from its score.

The old checker reproduces all $1{,}426$ archived first-completion grades; neither checker leaves an unresolved cell. Haiku passes $162/713$ questions under each checker. Sonnet passes $370/713$ under the old checker and $374/713$ under the revised checker: four scores change from failure to success and none in the other direction. Sonnet remains ahead in this two-model, first-completion comparison. One independent recomputation checks the input hashes, terminal logs, all grades and the unchanged ordering.

The separately specified ten-completion replay was incomplete under its registered $7{,}200$-second per-checker limit. The old and revised checkers completed 681 and 678 model–question rows. Keeping the original $14{,}260$ paired completion positions, $7{,}480$ pairs remain unknown. Among positions observed under both checkers, 43 move from failure to success and none move in the other direction. Three old-checker grades outside the first-completion endpoint also differ from the archived flags. With the unobserved positions retained in worst-case score intervals, the two-model ordering is unknown under both checkers. These partial counts cannot establish the historical ten-completion scores or a ranking change. A separate recomputation checks the fixed denominator, observed transitions and unknowns. The registered runner has no selective or continuation action; the unknown pairs are not free labels. Any selective endpoint needs prespecified actions, cache, failure handling and full costs, with equal-information, equal-risk and equal-coverage baselines. First- and ten-completion timings compare different targets, not verification savings. All three archive disagreements are non-first Haiku completions of one question. The archive records `TimeoutException` and a failing flag for each; both local checker versions return a passing flag. The historical execution environment was not reproduced. These replays neither correct the archived grades nor identify the checker revision as their cause.

This is one maintenance event and a fixed first-completion statistic, not the official average over ten samples, a reranking of the published leaderboard, or a test of semantic program correctness. The Lite release prunes or samples tests relative to the original problems. We provide hashes, derived counts and source retrieval instructions, but do not redistribute the archived programs or tests because the dataset card's abbreviated licence tag does not resolve the applicable variant.

W Related sensitivity and reuse methods

Fixed-output reevaluation and scoring-induced reranking are established phenomena (Table 31). We compare historical key, support and scorer operations on paired items and ask when a model comparison can be reused. Our bounded-edit rules hold within their stated classes; matched baselines show no empirical advantage over simpler rules.

Micro-benchmark pairwise reliability makes random subsets competitive at sufficient sizes (Yauney et al., 2026); task subsets preserve rankings under conditional bounds (Gusev & Zaytsev, 2026). Group-sequential evaluation stops fixed-protocol tests, sometimes using published predictions (Arviv et al., 2026). Exact shared-label winner, tolerance and regret certificates with query-order baselines exist for fixed weighted binary losses, including selective prediction (Kuboyama, 2026); active weak/strong-rater allocation minimizes mean-estimation error under an expected-cost budget (Angelopoulos et al., 2025). These works do not test published old/new scoring contracts with failed execution. Our study has not shown matched-risk, matched-information, full-cost savings.

SQL studies revise questions, evidence, references and databases (Jin et al., 2026a;b); our full-support fixed-output factorial separates reference and database effects, retains failures and lacks complete consumed prompts.

Table 31: Closest precedents and the narrower question examined here. Fixed outputs, rank changes or a stable champion are not priority claims.

Work	Established overlap	Comparison in this paper
Jo et al. (2025)	Answer extraction changes rankings; its post-processing conditions include fixed responses. Instructed formatting and answer regeneration are separate interventions.	Historical scorer behavior is reconstructed, then key, support and scorer changes are measured separately.
Kydlicek et al. (2025)	Replacing the operational mathematical scorer changes leaderboard scores and ranks while leading positions can stay fixed.	Scorer changes are contrasted with key-and-support revisions, without treating agreement with a parser as human semantic truth.
Panthi & Abdelfattah (2026)	Fixed memory-retrieval traces are rescored against changed credit targets, with shared-query controls, transfer, semantic auditing and mitigation checks.	Historical operations are placed in common sampling units; conditional edit budgets are tested against equal-retention and same-information baselines.

Measurement background. Label, ambiguity, grading and detection revisions are distinct interventions (Northcutt et al., 2021; Gema et al., 2025; Vendrow et al., 2025; Seo et al., 2025; Neuhaus & Hein, 2025; Truong et al., 2025; Penquitt et al., 2026). Paired uncertainty, variance and reproducibility motivate our sampling scale (Miller, 2024; Madaan et al., 2024; Perlitz et al., 2024; Biderman et al., 2024). Vintage-indexed data, revised outcomes and omission sensitivity have established treatments (Croushore & Stark, 2001; Diebold & Mariano, 1995; Clark & McCracken, 2009; Kuang, 2026; Broderick et al., 2020; Huang et al., 2026). Real-time predictive-ability tests address sampling distributions under revised data; our identities describe fixed outputs under a specified finite revision. Judge bias, drift, human-anchored attribution and verifier aggregation (Zheng et al., 2023; Dubois et al., 2024; Wang et al., 2023; Panickssery et al., 2024; Sunkavalli, 2026; Li, 2026; Krsteski et al., 2026) are not isolated by our protocol contrast.

Test equating links scores across forms (Kolen & Brennan, 2014); fixed-parameter calibration extends an IRT scale using shared anchors as benchmarks enter a suite (Habba et al., 2026). Neither a linked scale nor a stable fitted ability establishes that a changed answer key measures an unchanged target. Feng et al. (2026) use human-calibrated judge error rates for hypothesis testing; our conditional revision classes neither estimate those error rates nor justify their transfer between versions. Liu et al. (2026) change execution oracles for conformal SQL abstention, hold candidate pools fixed, and also perform same-SQL and expert checks. Some conditions select different final SQL; our historical factorial keeps each submitted SQL fixed. Incremental retrieval pooling reuses past judgments while holding old labels and ranked lists fixed, and separately examines judge sensitivity (Nelson et al., 2026). Thus historical reuse, scorer sensitivity and efficient selection each have direct precedents; their combination must still be tested under a common decision and cost contract.

X REFERENCE-ACTION REVISIONS IN A TOOL-USE ARCHIVE

WorkBench evaluates workplace tasks by executing tool calls and comparing the resulting database state with a reference outcome (Styles et al., 2024). We evaluate archived outputs against two reference-action targets while holding the current evaluator, tools and initial state fixed. This isolates a target effect in that execution environment, not the effect of the entire maintenance release or a replay of unknown historical tool versions.

Source and support. We pin the public WorkBench repository at commit `49c7dfd00c03d384ec59ea57374f50b766aa5613`. The source contains 24 all-tools model archives, each covering six domains and 690 tasks. We compare the archived target named `v2026-05-19` with the current `v2` target. The former is a version label, not evidence of a May publication date: the relevant corrections and archives were committed in June. Identity is domain plus the SHA-256 of exact task text, never a cross-version row number. This gives 679

shared tasks: 9 changed reference-action lists and 670 unchanged controls. The other 11 tasks are outside the fixed-text comparison. Two additional model archives with domain-filtered tools are not pooled with the all-tools roster. Model names retain upstream labels rather than independently authenticated service-version identities.

All matched tasks retain their template and domain metadata. The source version resolved for a run and the target used to evaluate it are recorded separately. Some source versions are explicit in the sidecar; others are the current loader's default, validated against the exact task set. Neither establishes the historical tool commit. The outputs are not uniformly known to precede maintenance, and this exposed archive is a descriptive development event. The unchanged-text `v1`–`v2` intersection has no action-target changes and is not the intervention studied here.

Native execution and calibration. Both endpoints use the official AST parser and fixed tool whitelist in an isolated Python environment. Official dispatch success, not human semantic validity, defines executable reference actions. The archived error field is preserved: it can veto final correctness even when the executed state matches. Unwanted side effects use state correctness before this error veto; exact match compares action lists rather than final-state equivalence. Model-action errors retain the official behavior, while an outer timeout or unresolved execution would remain unknown rather than receive an invented incorrect label.

The 679 old targets and all 690 current targets execute successfully. Six unchanged controls, one per domain, give identical scores under both targets on three repetitions. Recomputing all 16,560 current-target positions matches the pinned public release scores on all seven Boolean fields and the original error. We reuse those current-target scores by exact item identity and execute every old-target position, obtaining all 16,296 pairs. No model or judge is called.

Table 32: Changed judgments under fixed current execution. All changes occur on the 216 model–task positions with changed reference actions; the 16,080 control positions are unchanged.

Judgment	False to true	True to false	Total changed
Task success	145	34	179
Unwanted side effects	31	47	78
Exact action match	144	33	177

Scores and decisions. Table 33 reports all models on the same 679 tasks. Score-count changes range from -2 to 9; a reference revision does not improve every model. Four of the 276 unordered comparisons strictly reverse, and two old ties become strict comparisons. No new tie appears. The champion and the top-three and top-five sets are unchanged. Thus an unchanged selected model does not imply that all reusable comparisons, or all task-success judgments, are unchanged. These comparisons share models and tasks; they are not 276 independent revision events or the published 690-task leaderboard.

Cost and independent verification. The smoke-based runtime estimate exceeded the initial 1,200-second cap. Before any full paired outcome was read, a separate continuation retained that stop and set a 2,100-second cap using the measured estimate. The complete calibration and paired execution took 1,239.15 wall seconds, including reference execution within the native scorer. Preparation and independent verification are additional costs, not zero. This is measured execution cost, not a saving over a verification policy or a full reevaluation baseline. An independent implementation checked all row identities, aggregates and comparisons, and independently executed 90 endpoints from a frozen 45-pair subset covering every changed target and all six domains. All 630 checked Boolean values matched. This does not claim a second complete native replay. The supplement provides every paired score, hashes, a numerical replay and an optional offline native reproduction from the pinned upstream sources, without redistributing raw task or response text.

The event supports reference-target sensitivity in a second execution-based domain. It neither reproduces the arithmetic panel's cross-scorer contrast nor establishes future revision coverage, historical-tool invariance or a verification-policy advantage.

Table 33: Successful tasks out of 679 for every all-tools archive. Names are the upstream model labels. Both columns use one fixed current scorer, tool implementation and initial state.

Model label	Old target	New target	Change
claude-fable-5	654	663	9
claude-haiku-4.5	503	505	2
claude-opus-4.8	647	653	6
claude-sonnet-4.6	592	598	6
deepseek-v4-pro	568	570	2
gemini-3.1-pro	641	648	7
gemini-3.5-flash	615	624	9
glm-4.6	518	523	5
gpt-3.5	210	211	1
gpt-4-turbo	408	414	6
gpt-4.1	491	496	5
gpt-4o	460	466	6
gpt-5	568	571	3
gpt-5.1	409	412	3
gpt-5.2	473	478	5
gpt-5.4	525	531	6
gpt-5.4-mini	392	395	3
gpt-5.4-nano	327	327	0
gpt-5.5	636	645	9
kimi-k2.6	596	596	0
mistral-medium-3-5	409	413	4
mistral-small-2603	264	262	−2
o3	517	525	8
qwen-3.5-flash	469	477	8

Table 34: All strict reversals. Margins are first minus second model, in successful-task counts; divide by 679 for accuracy differences.

First model	Second model	Old margin	New margin
claude-sonnet-4.6	kimi-k2.6	−4	2
glm-4.6	o3	1	−2
gpt-4-turbo	gpt-5.1	−1	2
gpt-4-turbo	mistral-medium-3-5	−1	1