SELECTION IS NOT VALIDATION: AUDITING PROXY-TO-BEHAVIOR REVERSALS IN LANGUAGE-MODEL DATA CURATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-selection methods for language-model training are often validated with the same proxy used to construct the training set. This creates a circular evaluation shortcut: a selector can improve its target score without improving the trained model. We introduce a staged claim-advancement audit that treats selection scores as hypotheses rather than outcomes. It specifies random and credible simple-baseline comparisons, audits training parity, and organizes source-clustered uncertainty, paired reference-loss and generation analyses, retention checks, and mechanism diagnostics, while marking missing evidence explicitly. Unlike a pass/fail checklist, the audit maps each result to the strongest claim supported by its evidence and identifies the next experiment required to advance that claim. In linked discovery cases in mathematical pretraining and instruction tuning, stored evaluation records reveal source-dependent uncertainty, aggregation-dependent likelihood estimates, and disagreements between reference likelihood and decoded reference overlap on shared example IDs. These paired analyses and a single-annotator, conflict-enriched 20-pair calibration provide bounded diagnostic evidence; useful fitted-model results are retained even when broader claims remain unresolved. The resulting evidence map turns proxy optimization into an actionable validation program: reject unsupported downstream claims, qualify surface-specific gains, or advance only those effects that survive independent behavioral and mechanism tests.

1 INTRODUCTION

Data curation changes a training distribution before it changes a trained model. An increase in a selection score, a decrease in reference negative log-likelihood (NLL), and an improvement in generated answers are therefore different empirical claims. Our contribution is an operational procedure for deciding which of these claims a particular collection of evidence supports, and which additional measurement is needed to advance it.

The mismatch itself is established prior knowledge. LESS discusses why validation loss need not track instruction-tuning task performance (Xia et al., 2024); NICE selects for non-differentiable target metrics (Wang et al., 2025); and a large comparison of self-scoring methods shows why random selection is a serious baseline under its tested settings (Xia et al., 2025). We do not introduce a new selector or claim the first observation of proxy mismatch. We make evidence provenance, comparison units, metric definitions, and advancement rules explicit in one case-level ledger.

Evidence scope of this revision. The inspectable empirical evidence consists of three discovery pretraining contrasts, four matched instruction-tuning comparisons, and a small diagnostic human calibration. Earlier reported scale, placebo, and Mistral studies are retained in Appendix E, with unresolved provenance and required source records. Their available ledger is linked to a synthetic-data generator; it does not establish completed training or evaluation. We therefore do not use it to support empirical confirmation.

Three observations remain testable from stored evaluation rows. First, raw queryability selection has a small favorable NLL point estimate at 70M, but its interval crosses zero when resampling

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.

Stage	Required observation	Decision and next step
1. Static validity	Intended property changes; subset IDs, nondegeneracy, and simple confounds checked.	Permit a data-composition claim; otherwise repair selection.
2. Training parity	Same pool, initializer, protocol, and measured budget vector; random and credible simple baselines.	Permit a controlled recipe contrast only for matched dimensions; measure missing exposure.
3. Random uncertainty	Distinct selection/training seeds; paired run contrasts and absolute baselines.	Resolve variation at the declared unit; otherwise keep a fitted-run description.
4. Reference utility	Target and retention loss with named weighting and complete references.	Permit only the named loss claim; inspect aggregation and retention.
5. Generated behavior	Saved outputs on the same IDs, decoding settings, task metric and failure counts.	Permit a metric-specific effect; general quality needs independent evidence.
6. Matched uncertainty	Paired prompt/source intervals, strict signs, ties, missingness and sensitivity.	Resolve the stated estimand conditional on fitted models; change unit if dependence remains.
7. Mechanism	Score–benefit test and a discriminating control using a common evaluation protocol.	Permit mechanism consistency only within tested confounds; run direct true–shuffled contrast.
8. Confirmation	Independent seeds, domains or models; representative blinded quality assessment for quality claims.	Broaden only along observed dimensions; otherwise preserve the narrower claim.

Table 1: The eight-stage audit. Evidence completeness and scientific success are separate fields; passing a file-integrity check does not pass a stage.

source documents. Second, the RankTune comparison changes sign between prompt-macro and token-weighted completion NLL. Third, reference loss and ROUGE-L can disagree on identical evaluation prompts; exact metric ties and empty responses substantially affect the interpretation. ROUGE-L measures reference overlap, not general answer quality. Human labels on deliberately selected conflicts offer only a bounded diagnostic check, not population-level quality confirmation.

2 An Operational Claim-Advancement Audit

2.1 Objects, information, and budgets

Let $D = \{z_i\}$ be a candidate training pool, V a target specification or development set available before selection, and E a held-out evaluation set. A selector $g_\theta(D, V; b, \eta)$ returns a subset S under budget b and selection seed η. Its scores may use text features or a specified proxy model; they must not use post-training outcomes on E. We record this information boundary explicitly rather than assume that a target-conditioned score is intrinsically a validation result. Training maps $M = A(M_0, S; h, \xi)$, where M_0 is the initialization, h contains the optimization and tokenization settings, and ξ is the training seed. The baseline B is selected from the same candidate pool.

Budget is a vector, not an interchangeable scalar: $b = (n_{\text{examples}}, n_{\text{nonpad}}, n_{\text{updates}})$. Equal example counts or updates do not imply equal non-padding exposure. Matching the initializer, optimizer, stopping rule, truncation, and evaluation surface is also necessary for an unconfounded recipe contrast. Selection and training seeds are distinct sources of randomness. A paired test-set bootstrap conditional on one fitted model does not estimate either source.

2.2 Eight stages and explicit decisions

Table 1 defines the same eight stages used throughout the paper and artifact. Each claim record includes an evidence status (stored rows, aggregate only, or missing), a statistical status, the strongest permitted wording, and the next required observation. An unavailable stage is recorded as unresolved. A later diagnostic can describe a fitted model even when parity is incomplete, but cannot repair that gap for a causal or seed-general claim.

2.3 ESTIMANDS, DIRECTION, AND PRACTICAL RELEVANCE

All deltas in this revision are candidate minus baseline. For prompt i, $\Delta\ell_i = \ell_{S,i} - \ell_{B,i}$ and $\Delta r_i = r_{S,i} - r_{B,i}$; negative NLL and positive ROUGE-L are favorable. Retention NLL uses the same convention, so a positive retention delta is adverse. ROUGE-L is reported on $[0, 1]$; -0.017 is about -1.7 points. Prompt-macro and token-weighted NLL are respectively

$$\Delta L_{\text{macro}} = \frac{1}{n}\sum_i (\ell_{S,i} - \ell_{B,i}), \qquad \Delta L_{\text{token}} = \frac{\sum_i t_{S,i}\ell_{S,i}}{\sum_i t_{S,i}} - \frac{\sum_i t_{B,i}\ell_{B,i}}{\sum_i t_{B,i}}, \tag{1}$$

where t counts scored completion tokens. Common reference and tokenization checks precede interpreting these as a paired contrast. Neither weighting is universally preferable: each defines a different target quantity.

Define an oriented utility contrast as $U = -\Delta\text{NLL}$ for a loss or $U = \Delta\text{ROUGE}$ for overlap. With interval $[a, b]$ for U, practical advancement requires $a > \epsilon$, where $\epsilon \geq 0$ is an application-defined minimum useful effect fixed before a future confirmation. Intervals excluding zero but not ϵ resolve direction only. No defensible prespecified practical threshold is documented for these retrospective discovery cases; we report effect sizes and uncertainty without retroactively declaring one. A CI crossing zero is inconclusive, not evidence of equivalence or disappearance of the point estimate. Multiple related contrasts are exploratory; unadjusted 95% intervals are not a family-wise guarantee.

We distinguish an intended proxy change, a reference-loss effect, a reference-overlap reversal, a retention tradeoff, and an independently supported answer-quality difference. "Quality reversal" is reserved for the last category. A mechanism claim additionally requires a direct intervention that separates the proposed mechanism from shared recipe changes. A true–shuffled contrast is $\Delta\Delta = (S - B)_{\text{true}} - (S - B)_{\text{shuffled}}$ under a common evaluation protocol; subtracting unrelated published means is insufficient.

3 DISCOVERY CASES AND EVALUATION PROTOCOL

3.1 PRETRAINING AND INSTRUCTION TUNING

The pretraining cases use an OpenWebMath sample (Paster et al., 2024) and Pythia-70M-deduped (Biderman et al., 2023). They compare raw Queryable Tokens (QT) with uniform sampling (P1), intent-balanced QT with its raw-QT parent (P2), and frequency-capped QT with its raw-QT parent (P3). Each has one training seed, a nominal 10M-token subset budget, and 611 updates. The common validation map contains 3,894 chunks from 1,079 nonempty source documents. P2 uses a separate parent evaluation; substituting the P1 parent silently changes its estimand. These related variants do not establish independent replications or comparison with a mature external selector. An existing dedup-only 70M run supplies an additional simple comparator; its recovered contrasts appear in Appendix C, not as new training.

Instruction-tuning cases use Qwen2.5-0.5B (Qwen Team, 2024), LoRA (Hu et al., 2022), and fixed 5,000-example subsets from Alpaca- and Dolly-derived pools (Taori et al., 2023; Databricks, 2023). S1 compares relevance–diversity RankTune-MMR with random selection and includes a generation-rescue variant. S2 uses capped exposure correction, and S3 a retention-risk-penalized selector. The executable selector definitions, information boundaries and recovered configurations are in Appendix B. No LESS or NICE training run is present in the audited evidence, and none of the local selectors is relabeled as one. The target is a locally held-out, summarization-labeled subset of the training data sources, not an external task benchmark. Target-development data guide relevance, while target-test text and references enter duplicate-risk control. Thus test content was not completely hidden from selection; this visibility does not establish outcome-based tuning but limits an untouched-test claim.

Parity qualification. These designs match nominal examples or iterations, not all realized token exposure. In P3, frequency capping also changes artifact/near-duplicate filtering and consumes about 1.32% fewer non-padding tokens than its parent. Thus even a resolved held-out loss difference is a recipe contrast, not an isolated causal effect of frequency. SFT logs report unequal supervised completion-token counts: 24,452 versus 39,699 for random/MMR, 11,804 versus 13,343 for ran-

Case	Baseline	Candidate	ΔNLL	Unit	95% CI
P1	3.535936	3.534082	-0.001854	Chunk	$[-0.002821, -0.000882]$
				Source	$[-0.006487, +0.002271]$
P2	3.534622	3.536587	$+0.001965$	Chunk	$[+0.000738, +0.003202]$
				Source	$[-0.002067, +0.005862]$
P3	3.534082	3.538101	$+0.004019$	Chunk	$[+0.002530, +0.005597]$
				Source	$[+0.001526, +0.006856]$

Table 2: Discovery Pythia-70M comparisons. Each contrast uses 3,894 chunks from 1,079 sources, one fitted training seed and 1,910,597 scored validation tokens. Deltas are candidate minus baseline; lower is better. Source and chunk intervals are independently recomputed with 10,000 draws.

dom/exposure, and 13,604 versus 14,856 for random/guarded. These counters are not total prompt-plus-completion exposure. With gradient accumulation, reported MLX iterations are microbatches, not optimizer updates.

3.2 Stored-output statistical reproduction

The anonymous artifact contains paired pretraining loss CSVs and a source map, plus 998 matched completion-complete SFT comparison rows. These comprise 200 MMR, 200 rescue, 498 exposure and 100 guarded pairs; the two S1 arms reuse evaluation prompts and random outputs. Across all comparisons there are 498 distinct prompt IDs: S1's 200 prompts are contained in S2, and S3's 100 are contained in S1 and S2, with matching prompt/reference content. They are 998 comparison rows, not 998 independent tasks or model runs. Two exposure records have incomplete references and are excluded from paired NLL; their saved generations remain eligible for the separately labeled 500-prompt generation sensitivity.

We resample paired chunks or source documents for pretraining, and paired prompt records within each SFT comparison. A document draw carries all of its chunks and tokens, preserving a token-weighted loss ratio rather than averaging document means. All primary intervals in this revision use 10,000 percentile bootstrap draws with fixed documented analysis seeds. These intervals condition on the fitted models and held-out sampling surface. They do not quantify training-seed uncertainty, leakage, annotation reliability, or model transfer. Appendix C gives implementation details and the distinction between recomputing stored-output statistics and rerunning model inference.

Empty responses are kept in the primary generation analysis and counted separately; a nonempty-only slice changes the population and is sensitivity analysis. Strict quadrants use $\Delta\ell < 0/> 0$ and $\Delta r > 0/< 0$. Any exact zero is a tie, reported outside the four strict cells. In particular, NLL-better/ROUGE-equal is not NLL-better/ROUGE-worse. The appendix records empty transitions and their positions in this sign classification.

4 Results from Traceable Records

4.1 Source dependence changes uncertainty

Table 2 places the absolute baselines and both uncertainty units together. P1 changes NLL from 3.535936 to 3.534082, a relative reduction of about 0.0524%. Its chunk interval excludes zero, whereas its source interval does not. P2 has the same uncertainty distinction in the adverse direction. P3 remains adverse under source resampling, but the measured token and filtering differences prevent attribution to frequency capping alone. This is why the ledger records separate fields for statistical resolution and causal scope. Aggregate gain establishes no mechanism; the historical score–benefit diagnostics are retained in Appendix A with their evidence grade.

4.2 Aggregation and decoded overlap answer different questions

Table 3 uses a single delta convention. RankTune-MMR has unfavorable, unresolved macro NLL but favorable token-weighted NLL, alongside lower ROUGE-L. The rescue arm has unresolved macro NLL, favorable token-weighted NLL, and lower ROUGE-L. Exposure correction has worse comple-

Comparison	n	Metric	Candidate − baseline	95% CI
S1 MMR	200	Macro NLL	$+0.004747$	$[-0.002008, +0.012146]$
		Token NLL	-0.008068	$[-0.011367, -0.004730]$
		ROUGE-L	-0.017046	$[-0.033474, -0.002970]$
S1 rescue	200	Macro NLL	-0.000255	$[-0.006272, +0.006280]$
		Token NLL	-0.007877	$[-0.011079, -0.004655]$
		ROUGE-L	-0.011934	$[-0.024401, -0.000613]$
S2 exposure	498	Macro NLL	$+0.032990$	$[+0.027516, +0.039169]$
		Token NLL	$+0.015878$	$[+0.013404, +0.018553]$
		ROUGE-L	-0.009080	$[-0.019370, +0.000807]$
S3 guarded	100	Macro NLL	-0.004316	$[-0.007485, -0.000572]$
		Token NLL	-0.003858	$[-0.005743, -0.002096]$
		ROUGE-L	-0.005511	$[-0.012336, +0.000308]$

Table 3: Matched-content discovery SFT comparisons, conditional on the saved models and evaluation protocols. NLL and ROUGE-L have opposite favorable directions. NLL chat formatting differs from generation formatting; these are not matched-template causal contrasts.

Comparison	n	$\ell\downarrow r\uparrow$	$\ell\downarrow r\downarrow$	$\ell\uparrow r\uparrow$	$\ell\uparrow r\downarrow$	Ties	Empty B/S
S1 MMR	200	48	52	22	31	47	12/22
S1 rescue	200	44	52	22	33	49	12/15
S2 exposure	498	42	48	147	132	129	20/48
S3 guarded	100	17	18	5	11	49	23/23

Table 4: Four strict sign cells plus all exact-tie cells. Strict cells and ties sum to n. Empty B/S gives counts for baseline/selector (including shared empties); they are not additional rows. Detailed nine-cell and empty-transition counts appear in the appendix.

tion NLL under both weightings; its average ROUGE-L change is negative but its paired interval crosses zero. Guarded selection improves both completion-NLL estimands, while the ROUGE-L interval also crosses zero. The latter two ROUGE results cannot be described as statistically resolved overlap deterioration.

The historical MLX corpus test loss and matched completion-only rescoring are different evaluation protocols. We do not combine their numerical values or switch signs according to whether a column is called "improvement." The artifact exposes absolute losses, scored token counts and paired differences to make this distinction inspectable. Completion likelihood is conditional on the reference; free-generation overlap is conditional on the decoding policy. Agreement is an empirical question, not a property of the NLL objective. There is also an evaluation-format confound: the saved generations use a literal instruction/input/response prompt, whereas the completion-NLL scorer uses Qwen's chat template. Same IDs therefore establish matched content, not an identical tokenized context. The observed disagreement is conditional on these two protocols; a matched-template rerun is required to isolate the teacher-forcing versus free-generation distinction.

4.3 Strict quadrants, ties, and empty responses

Table 4 keeps ties outside strict cells. For MMR, the strict conflict count is 52/200 (26%); the older nonpositive-ROUGE bucket is 69/200 because it includes 17 exact ROUGE ties. For guarded, 18/100 are strict conflicts, whereas 49/100 are NLL-better with nonpositive ROUGE; 31 of these have unchanged ROUGE. A reference-loss improvement on an identical decoded response is not evidence of a quality loss.

Exposure correction increases empty responses from 20/498 to 48/498 in the matched set. That concrete failure diagnostic is more directly interpretable than its unresolved average ROUGE delta. MMR increases empties from 12/200 to 22/200. Guarded has 23 shared-empty pairs and no one-arm-only empty transition; 46/100 decoded response pairs are identical. These counts describe the saved decoding setup and do not establish failure rates under alternative prompts, sampling temperatures or stop rules.

Case	Currently supportable statement	Unresolved claim and minimum next test
P1	Small favorable fitted-run NLL difference; source CI crosses zero.	Seed-general benefit: repeat selection/training with matched exposure and source evaluation.
P2	Adverse point estimate; source-level uncertainty remains.	Intent-balancing effect: preserve parent, measure exposure, repeat seeds.
P3	Adverse source-resolved recipe contrast.	Frequency mechanism: isolate filtering and token exposure, then true–shuffled control.
S1	Macro/token estimands disagree; token NLL improves and ROUGE decreases.	General answer quality: random prompts, independent blinded raters, task-valid metrics.
S2	Completion NLL worsens; saved empty outputs increase.	Cross-domain harm: rerun matched generation and training seeds with valid domain metrics.
S3	Completion NLL improves; mean ROUGE remains unresolved.	Retention/quality benefit: matched retention rows, representative quality evaluation.

Table 5: Claim boundaries for the six linked discovery cases. None is a fully confirmed seed-general utility claim. Positive loss evidence is retained without bypassing an earlier parity or randomness gap.

The previously reported 7B counts 517 and 454 do not come from these discovery records. Their four cells sum to 2,000, but no matched 7B prompt outputs have been supplied and the associated generator samples preset categorical probabilities. Their implied percentages, 25.85% and 22.70%, are arithmetic properties of a historical ledger, not new behavioral measurements (Appendix E).

4.4 Human calibration has a diagnostic scope

The recorded calibration consists of 20 deliberately selected comparisons: 12 S1 and eight S3, including 16 strict-conflict cases and four tie/identical controls. One unpaid annotator, described by the author as a Linguistics Olympiad participant, labeled all pairs. This description establishes neither age, author relationship, consent nor an ethics determination. The rubric requests task fulfillment, coverage, fluency, unsupported content, truncation or emptiness, overall preference, and confidence. Responses were presented with A/B identifiers; the procedure and available masking records are described in Appendix D.

The recorded aggregate reports 20 usable pairs, 140 completed required dimensions and median confidence 3/3. Three identical-output controls were ties; a fourth non-identical ROUGE-tie control favored the baseline. Among the 16 strict conflicts, baseline responses were preferred 11 times and selector responses five times. Even treating these as 16 independent fair-coin trials, the two-sided exact sign-test p is approximately 0.210; the selected sample does not warrant a population inference in any case. There is no second annotator, inter-annotator agreement estimate or random-sample validation.

These labels show that some deliberately identified automatic disagreements can be examined by a human. They do not establish that ROUGE-L is a sufficient quality metric or that the selector worsens answers on the full distribution. A described calibration gate is a workflow rule; its existence and temporal freeze are separate questions. We do not label the discovery study preregistered or claim ethics approval or consent without records.

5 Claim Decisions, Missing Evidence, and Next Tests

The procedure produces useful partial outcomes (Table 5). For example, the guarded fitted model has a resolved favorable completion-NLL contrast under both aggregations, while its generated-overlap effect remains uncertain. The ledger can preserve this result without claiming broad quality improvement. In P3, changing the uncertainty unit does not remove the adverse contrast; changing the causal interpretation is still necessary because the recipes differ along more than one dimension. These are constructive decisions, not a rule that rejects every selected subset.

The regularity of the earlier scale results requires a provenance check before an explanatory narrative. Their available generator sets effect sizes in constants and adds random Gaussian perturbations;

model names and seed labels are not evidence that a model ran. C2 checkpoint identifiers are constructed strings, and their matching check does not load or authenticate weights. The eight SFT scale–domain cells therefore cannot be treated as independent confirmations or used to explain why discovery uncertainty became smaller. Appendix E preserves the reported designs and the exact missing records. Actual runs, if available elsewhere, require an independent manifest, commands, logs, checkpoints and per-example measurements before reassessment.

The minimum next test repeats one discovery contrast across actual training seeds with measured exposure and a common prompt template. True and shuffled scores require a shared budget, initialization and evaluation protocol. A mature-selector comparison then tests whether the audit changes conclusions beyond custom recipes. Scale cannot repair a missing baseline or uncertainty unit.

6 Related Work and Scope of the Contribution

LESS uses a small target validation set and gradient-based influence for targeted instruction tuning (Xia et al., 2024). Its empirical settings and selection objective differ from our simple relevance and risk recipes. NICE targets non-differentiable metrics through influence estimation (Wang et al., 2025); it addresses metric alignment algorithmically. Random Selection Is Almost All You Need compares self-scoring data selection for LLM fine-tuning (Xia et al., 2025); its results motivate a strong random baseline within that scope, not a theorem that all curation methods fail. These works already undermine an automatic inference from selection score or validation loss to downstream utility. We make no performance comparison against them without an actual run.

The audit adds an explicit evidence status and next experiment to each information boundary, estimand and uncertainty unit. Its reconstructable case series demonstrates source dependence, aggregation disagreement, ties and decoding failures, not a universal metric or selector ranking. Whether these decision rules improve research practice beyond existing recommendations requires a separate evaluation; this case series does not measure that impact.

7 Limitations and Conclusion

The retrospective cases share data, models and baselines. Evaluation bootstraps do not cover training randomness, and unequal exposure, coupled filtering and different prompt templates restrict causal attribution. ROUGE-L is a lexical summary metric (Lin, 2004); the single-annotator, conflict-enriched calibration cannot establish general quality, prevalence or reliability. Earlier scale and cross-family results remain blocked by missing independently traceable run evidence.

A selection score supports a statement about the selected data. Advancing to a claim about a trained model requires a named outcome, a justified comparison and uncertainty unit, and inspectable evidence. In the recoverable discovery records, source dependence, aggregation choice and generation diagnostics each change the strongest supportable conclusion. The audit makes those boundaries and next tests explicit while retaining useful, narrowly defined observations.

Ethics Statement

The human evidence is a 20-pair, single-annotator diagnostic calibration. The known author-provided description is an unpaid Linguistics Olympiad participant. Age, relationship to the authors, consent documentation and any institutional review or exemption determination remain unverified in this revision. We make no affirmative claims about them. These disclosures require author resolution before public release of participant-level material; the anonymous artifact therefore contains the de-identified aggregate and clearly states its scope. Public training and evaluation data remain subject to their original licenses.

Reproducibility Statement

The anonymous artifact recomputes stored-output statistics with recorded bootstrap settings and separates row-level, aggregate-only and missing evidence. It does not retrain models, regenerate pre-

dictions or authenticate checkpoints. Appendix B documents recovered protocols and gaps; hashes establish file integrity only.

AI USE STATEMENT

AI tools assisted writing, coding, numerical audit, literature/policy lookup and artifact preparation. Historical ledgers produced by a synthetic-data script are excluded from empirical support. The authors remain responsible for all claims, numbers, references and released files. AI is not an author, human annotator or independent experimental evidence.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Databricks. databricks-dolly-15k. https://huggingface.co/datasets/databricks/databricks-dolly-15k, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jKHmjlpViu.

Qwen Team. Qwen2.5 technical report, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Pang Wei Koh, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. NICE data selection for instruction tuning in LLMs with non-differentiable evaluation metric. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63662–63689, 2025. URL https://proceedings.mlr.press/v267/wang25bm.html.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2698–2711, 2025. doi: 10.18653/v1/2025.findings-emnlp.146.

A Evidence Map and Reproduction Coverage

The accompanying artifact distinguishes three operations: input-integrity checking, recomputation of statistics from stored evaluation outputs, and regeneration of model outputs. Only the first two are performed in this revision. Its `claim_evidence.json` links each core claim to input files, sampling units, analysis code and paper tables. The JSON report records the actual bootstrap settings and hashes. A successful archive or hash check is never labeled raw experimental reproduction.

Claim / table	Available input and unit	Verified scope / missing evidence
P1–P3; Table 2	Paired loss CSVs plus chunk-to-document map; 3,894 chunks, 1,079 documents per contrast.	Absolute token-weighted loss and paired evaluation bootstrap recomputed. One training seed; no training replay.
Simple comparator; Table 7	Additional dedup-only loss CSV; shared uniform and raw-QT arms.	Two post-hoc contrasts recomputed; not a new seed, independent case or strong baseline.
S1–S3; Table 3	`paired_rows.jsonl`; 200 MMR, 200 rescue, 498 exposure, 100 guarded rows.	Macro/token loss, ROUGE and paired intervals recomputed from stored columns. Fresh inference and identical prompting not established.
Quadrants; Table 4	Same paired rows with output token counts and identity flags.	Nine strict/tie cells and empty transitions recomputed. Exact underlying decoded text is a separate source layer.
Human; Section 4.4	Twenty local label records and private A/B mapping; released aggregate only.	Local counts can be checked; public package cannot reconstruct labels or blinding. No random sample, second human or ethics record.
Static proxies	Stored case/proxy summaries and selector code.	Summary arithmetic and intended direction inspectable; full selection regeneration not performed.
Retention	Historical SFT aggregate evaluator summaries.	Small rounded deltas are descriptive only; matched retention evaluation and uncertainty are not supplied in the released statistics.
Scale / C1 / C2	Earlier numeric ledger; associated synthetic generator.	Historical records retained, empirical provenance blocked; no actual training-seed CI or transfer claim admitted.
Sample-size study	Analytic width scaling and preset concordance.	A planning illustration only, not observed bootstrap stability or replication.

The 998 SFT comparison rows cover 498 distinct prompt/reference contents, with S3's 100 contained in S1's 200 and S2's 498. S1 also shares the random baseline across MMR and rescue. They must not be pooled as independent samples. P1 and P3 reuse the raw-QT parent; P2 has a distinct parent evaluation. Reuse is recorded explicitly rather than counted as new replication. The revised supplement does not contain synthetic predictions, fabricated training logs, or the historical generated ledger as measured data. That ledger remains in the authors' local audit archive, and the associated experiment designs remain below.

The historical P1 case summary reports Spearman correlations of -0.0475 and -0.0385 between absolute loss reduction and QT score and queryability, respectively. These are retained as aggregate mechanism diagnostics, not recomputed by the released row-statistics script; the complete aligned score and pre-/post-training loss vectors are needed to reproduce that mechanism test. Their sign does not establish the success of the proposed mechanism.

B Recovered Protocols and Information Boundaries

The following settings are supported by saved discovery manifests and logs. Current implementation details are distinguished from recorded run settings; the original environment and repository revisions were not fully pinned.

Continual pretraining. All discovery models start from `EleutherAI/pythia-70m-deduped`. The trainer uses full-model AdamW, learning rate 5×10^{-5}, weight decay 0.01, warmup ratio 0.03, maximum sequence length 512, batch size 4, gradient accumulation 8, and training seed 1. It calculates 611 optimizer updates from a nominal 10M-token budget. It does not stop at a fixed count of non-padding tokens. The actual counters sum attention masks over consumed microbatches.

Run	Updates	Non-padding input tokens
P1 uniform	611	9,483,821
P1 raw-QT / P3 parent	611	9,496,068
P2 raw-QT (separate CPU run)	611	9,496,068
P2 intent-balanced	611	9,570,701
P3 frequency-capped	611	9,371,046

The P1/P3 runs use MPS; P2 uses CPU. P2's parent is a separate run even though its consumed-token count agrees. These are update-matched comparisons with unequal exposure, including 1.32% fewer tokens for frequency capping than its parent. They do not satisfy exact token parity.

Instruction tuning. Saved configurations identify the base model as `Qwen/Qwen2.5-0.5B`, LoRA rank 8, scale 20, dropout 0, eight adapted layers, Adam, batch size 1, accumulation 16, prompt masking, and maximum sequence length 768. Each selected set contains 5,000 examples. This is a selection budget; the bounded training schedule need not consume every selected example.

Comparison	Iterations	LR	Seed	Completion tokens, random / candidate
S1 RankTune	400	5×10^{-6}	13	24,452 / 39,699
S1 rescue	400	5×10^{-6}	13	24,452 / 32,812
S2 exposure	200	10^{-5}	0	11,804 / 13,343
S3 guarded	200	5×10^{-6}	13	13,604 / 14,856

The logged token counter counts masked completion positions, not total non-padding input exposure. The available MLX-LM implementation counts iterations as microbatches and updates only every 16 iterations; accordingly, iterations must not be described as optimizer steps. A historical runtime snapshot is needed to certify exact update and final-accumulation semantics. S2's original validation log contains NaN and truncation warnings; finite completion-NLL rescoring is a later, separate evaluation.

Selection rules. Raw-QT combines robustly standardized BM25 self-retrieval queryability, band-pass difficulty, inverse popularity, redundancy, boilerplate popularity, and phrase overlap. The trained subset uses weighted sampling without replacement with weight $0.2+0.8(0.05+q)$, selection seed 1, and a nominal 10M-token budget. It is not deterministic top-score selection. Intent-balanced QT reweights the QT prior by broad/narrow query balance, query-type entropy, and a validity factor penalizing exact or lead-overlapping pseudoqueries. The trained strict recipe targets 90/7.5/2.5% token quotas for balanced/broad-dominant/narrow-dominant buckets, with domain and SimHash-prefix caps of 6% and 8%. These differ from current non-strict defaults. Frequency-capped QT ranks by raw QT, excludes heuristic artifact risk at least 0.25, limits near-duplicate groups to four, and increases a SimHash-prefix cluster cap until the nominal budget fills; the recorded effective cap is eight.

RankTune greedily combines a pointwise relevance/quality/difficulty score with diminishing target-cluster coverage, maximum cosine-redundancy cost, and selected-source share. The source provides multiple coefficient configs; the saved selection summary does not bind a complete historical config

Comparison	ΔNLL	Source 95% CI
Dedup minus uniform	$+0.007036$	$[+0.003781, +0.010450]$
Raw QT minus dedup	-0.008891	$[-0.016740, -0.002148]$

Table 7: Recovered dedup-only simple comparator, evaluated post hoc on the same 3,894 chunks and 1,079 sources. One training seed; 10,000 paired source-bootstrap draws. These contrasts reuse discovery arms, and do not establish a strong-method ranking.

hash. Exposure correction uses a frequency-based *pseudo*-propensity derived from source, task, cluster, response style, and duplication, floors it at 0.05, caps inverse weights at 10, and selects quality-filtered examples under group coverage caps. These are not observed logging propensities, so the procedure does not establish an unbiased causal correction. Guarded selection combines relevance with heuristic quality, length/style and risk features, group caps, and coverage floors. Its risk score is not an independently validated retention predictor. Local influence and diversity proxies are not original LESS, NICE, MADS, or BIDS reproductions.

Target visibility and shared examples. The instruction target is a locally held-out, summarization-labeled subset of the Alpaca/Dolly pool rather than an external benchmark. Some task labels are heuristic. Relevance and target clusters use development examples; duplicate risk uses development and test texts, including references. Thus test text is visible for duplicate control before training. The discovery comparisons share source pools and underlying prompts, and are not independent domains.

Prompt-format confound. Stored generations use literal `Instruction:`, optional `Input:`, and terminal `Response:` fields. Completion-NLL rescoring instead applies the model's chat template to user/assistant records. The example IDs and references match, but the tokenized conditioning contexts do not. Consequently, the observed NLL–ROUGE disagreement describes two evaluation surfaces and does not isolate selector-induced quality degradation under a common prompt format. A matched-template evaluation is required for that stronger claim.

Saved generation manifests specify greedy decoding (temperature 0), maximum 128 new tokens, top-$p = 1$, top-$k = 0$, min-$p = 0$, batch size 1, seed 13, and non-dry-run execution. Generated strings are stripped; an empty output has zero remaining characters. Completion scoring left-trims sequences to 768 tokens and records dropped prompt and reference tokens separately. Current training code truncates on the right; this difference is another protocol limitation rather than evidence of matched truncation.

Blinded calibration. The fixed diagnostic queue has 20 pairs: 12 from S1 (six original RankTune, six rescue) and eight from S3, including 16 strict metric conflicts and four controls. One human annotator judged randomly assigned A/B sides without arm names. The instructions prioritized task fulfillment over reference overlap; Chinese translations were aids and English was authoritative for subtle judgments. A label-completeness manifest and hashes support file traceability, but do not constitute external preregistration or an independent record of freeze timing. A separate AI sensitivity pass is not a second human annotator and cannot establish inter-human reliability. The deliberately enriched sample does not estimate population preference or mismatch prevalence. The author reports that the annotator was an unpaid linguistics-Olympiad candidate; age, authorship relation, informed consent, and ethics review or exemption remain author-confirmation items. No claim of clearance is made.

C Statistical Definitions and Detailed Recomputed Results

C.1 Paired bootstrap and absolute baselines

An existing dedup-only run provides a simple comparator with absolute NLL 3.542973, versus 3.535936 for uniform and 3.534082 for raw-QT. Table 7 reports its post-hoc source-level contrasts. Deduplication is worse than uniform here, so it is not labeled an established strong baseline. All three use seed 1 and 611 updates; dedup consumes 9,483,504 non-padding tokens, compared with

Comparison	Metric	Random baseline	Candidate
S1 MMR	Macro NLL	1.797927	1.802674
	Token NLL	1.642527	1.634459
	ROUGE-L	0.281561	0.264515
S1 rescue	Macro NLL	1.797927	1.797672
	Token NLL	1.642527	1.634650
	ROUGE-L	0.281561	0.269627
S2 exposure	Macro NLL	1.843659	1.876649
	Token NLL	1.695926	1.711804
	ROUGE-L	0.289134	0.280054
S3 guarded	Macro NLL	1.858039	1.853723
	Token NLL	1.746329	1.742472
	ROUGE-L	0.232714	0.227203

Table 8: Absolute means recomputed from paired stored outputs. The S1 variants share the same random baseline. These are completion-only rescoring values, not historical MLX training-validation losses.

9,483,821 for uniform and 9,496,068 for raw-QT. These are additional contrasts of existing arms, not new seeds or independent cases.

For chunk j, the artifact records loss ℓ_{mj} and scored-token count t_j for fitted model m. All compared chunk IDs and token counts must match. For resampling unit u, form $d_u = \sum_{j \in u} t_j(\ell_{Sj} - \ell_{Bj})$ and $w_u = \sum_{j \in u} t_j$. Draw units with replacement and compute $\sum d_u / \sum w_u$ in each bootstrap draw. A chunk bootstrap sets $u = j$; a source bootstrap sets u to the full document. The latter samples documents equally while retaining within-document token weights. It is not an equally weighted average of document losses.

SFT macro NLL and ROUGE use the mean paired difference in each prompt draw. Token-weighted completion NLL uses stored loss sums divided by the corresponding completion-token sum in each draw. The scorer uses the same reference tokens for both arms. Primary intervals are the 2.5th and 97.5th percentiles of 10,000 draws with linear interpolation. PT uses analysis seeds 1 (chunk) and 20260716 (source); SFT uses 20260716 (macro), 20260719 (token) and 20260717 (ROUGE). These are analysis seeds, not new model training seeds. Earlier PT reports used fewer draws and slightly different percentile conventions; the audit records the resulting numerical changes without presenting them as new experimental observations.

Bootstrap draws are not independent training replications. We do not report a bootstrap tail fraction as a hypothesis-test p value. Three-seed Student-t intervals would apply to actual independent paired run differences, with $\bar{d} \pm 4.303 s_d/\sqrt{3}$ for $df = 2$. The historical synthetic scale ledger does not meet that input requirement. No true–shuffled paired interval is recomputed in the absence of genuine common-protocol outputs.

Table 9: All nine sign cells and zero-token output allocation. Better means lower NLL or higher ROUGE. Empty categories are disjoint subsets of the cell count.

Comparison	NLL/ROUGE	Count	Both empty	B-only	S-only
S1 MMR	better/better	48	0	0	0
S1 MMR	better/worse	52	0	0	5
S1 MMR	better/tie	17	7	0	0
S1 MMR	worse/better	22	0	0	0
S1 MMR	worse/worse	31	0	0	5
S1 MMR	worse/tie	29	5	0	0
S1 MMR	tie/better	0	0	0	0
S1 MMR	tie/worse	0	0	0	0
S1 MMR	tie/tie	1	0	0	0
S1 rescue	better/better	44	0	0	0
S1 rescue	better/worse	52	0	0	1

Comparison	NLL/ROUGE	Count	Both empty	B-only	S-only
S1 rescue	better/tie	25	8	0	0
S1 rescue	worse/better	22	0	0	0
S1 rescue	worse/worse	33	0	0	3
S1 rescue	worse/tie	21	3	0	0
S1 rescue	tie/better	1	0	1	0
S1 rescue	tie/worse	2	0	0	0
S1 rescue	tie/tie	0	0	0	0
S2 exposure	better/better	42	0	0	0
S2 exposure	better/worse	48	0	0	10
S2 exposure	better/tie	32	9	0	0
S2 exposure	worse/better	147	0	0	0
S2 exposure	worse/worse	132	0	0	17
S2 exposure	worse/tie	95	11	0	1
S2 exposure	tie/better	0	0	0	0
S2 exposure	tie/worse	1	0	0	0
S2 exposure	tie/tie	1	0	0	0
S3 guarded	better/better	17	0	0	0
S3 guarded	better/worse	18	0	0	0
S3 guarded	better/tie	31	16	0	0
S3 guarded	worse/better	5	0	0	0
S3 guarded	worse/worse	11	0	0	0
S3 guarded	worse/tie	17	7	0	0
S3 guarded	tie/better	0	0	0	0
S3 guarded	tie/worse	0	0	0	0
S3 guarded	tie/tie	1	0	0	0

C.2 Missing references, empty responses, and ties

The primary S2 comparison excludes two completion-truncated reference records, leaving 498. Excluding incomplete reference loss does not justify dropping their generations from every analysis. The historical generation-only 500-prompt ROUGE delta was -0.009041, versus approximately -0.009080 on the completion-complete 498; this is a separately identified stored summary. Nineteen of the 498 primary S2 prompts were prompt-truncated during scoring. The untruncated 479-row subset is a sensitivity estimand, not a replacement for the originally evaluated distribution. Neither removal estimates training variation.

Empty counts in the paired artifact use zero prediction tokens under the saved ROUGE evaluator. This operational definition differs from proving that a decoded string contains no characters in general. In the inspected original generations, zero-token flags and stripped empty strings agreed on every record; the public numeric rows alone cannot recreate that text-level check. Shared-empty outputs can still have different reference NLL, so they need not be NLL ties. The detailed table locates these outputs by strict NLL and ROUGE signs. Exact ROUGE ties are all retained. We do not add an undocumented floating tolerance or force sign-tie rows into one of four quadrants.

D Human Calibration: Records and Unresolved Disclosures

The queue contains a fixed, conflict-enriched set of 20 IDs, with A/B response order randomized using seed 20260716. The queue-building script separates public forms from the private arm mapping. Available instructions ask that the key remain hidden until labeling is complete; an accompanying frozen-label snapshot records the intended workflow. These records support the described procedure but do not independently prove the timing of every access or a prospective preregistration.

The described calibration gate asks for at least 18/20 usable pairs, at least 90% required-dimension completion, median confidence at least 2, identical controls called ties, and a consistent direction in at least three diagnostic strata. The aggregate meets these descriptive criteria, but this does not turn a selected batch into random-sample validation. The local label recount and released summary check have distinct coverage. No inter-human reliability is estimable; any AI-assisted sensitivity review is

Diagnostic stratum	Baseline	Selector	Tie	Total
S3 high conflict	4	2	0	6
S1 empty transition	2	0	0	2
S1 high conflict	3	2	0	5
S1 format constraint	2	1	0	3
Strict-conflict subtotal	11	5	0	16
Non-identical ROUGE-tie control	1	0	0	1
Identical-output controls	0	0	3	3

Table 10: Recorded single-annotator preferences. The four stratum counts do not represent four independent studies.

not a second human rater. The earlier 2,000-prompt simulated preference exercise is not participant evidence and is excluded from scientific support.

Before publication the authors must confirm the annotator's age category, relationship to the authors, recruitment and consent procedure, and applicable institutional review or exemption determination. An unpaid Olympiad background alone answers none of those questions. Participant-level records and the private arm key are not included in the release candidate. No inference is made that ethics review was unnecessary because the task concerned model outputs.

E Previously Reported Confirmation: Preserved, Not Verified

This section preserves the earlier experimental scope and reported numerical claims for author reconciliation. The entries are not empirical results of the present revision. The inspected generator constructs PT/SFT baselines and effects as constants plus Gaussian draws. It also constructs checkpoint IDs as strings and samples four-quadrant counts from preset probabilities. Numeric closure of those outputs cannot authenticate training or evaluation. If genuine runs exist elsewhere, they must be linked independently; the absence of those files here is an internal release blocker, not a conclusion about an author's intent.

Preserved study	Earlier reported design / ledger values	Evidence required before use
Pythia scale matrix	410M/1.4B/2.8B/6.9B; 4 policies $\times$ 3 seeds = 48 nominal units. Raw-QT target CI excluded zero at 3/4 scales; retention deltas about .00391/.00334/.00355/.00353.	Actual model IDs/revisions, 48 run manifests, measured exposure, checkpoints and target/retention rows.
Qwen2.5 scale matrix	1.5B/7B/32B/72B; 4 selectors $\times$ 3 seeds = 48 nominal adapters; two domains, 2,000 prompts each. Earlier all-cell ROUGE patterns were unusually regular.	Adapter files, exact base/instruct variant, training seeds, per-domain prompts and saved outputs; absolute baselines for every seed.
C1 placebo	6 pretraining and 9 SFT nominal new units; true seeds 17/42/137, shuffled/control seeds 201–203.	True and shuffled results under the same scoring/decoding protocol; independently justified pairing or unpaired run analysis for $\Delta\Delta$.
C2 Mistral	7B-v0.3 and Nemo-12B; 2 models $\times$ 3 arms $\times$ 3 seeds = 18 claimed units. PT raw-QT means $-.00286$ and $-.00380$.	Exact repository/revision and training lineage; PT/SFT checkpoint mapping with real weight files; outputs from each evaluation.

Preserved study	Earlier reported design / ledger values	Evidence required before use
7B quadrants	RankTune 454/517/368/661; exposure 188/737/246/829; guarded 491/346/414/749; each sums to 2,000.	Prompt IDs and joint losses/outputs; exact-tie policy and empty-output cell allocation. Generated categorical counts are insufficient.
C3 stability	Concordance 75.0/84.0/92.6/100% at n=250/500/1000/2000; analytic $1/\sqrt{n}$ widths.	Actual resampling over measured outputs; arbitrary concordance points cannot be justified by the width formula.
Large human exercise	2,000-prompt simulated preferences and illustrative agreement.	Genuine blinded labels, sampling and participant disclosures; no synthetic label may substitute.

The earlier quadrant order was NLL-better/ROUGE-better, NLL-better/ROUGE-worse, NLL-worse/ROUGE-better, NLL-worse/ROUGE-worse. The implied RankTune conflict rate is $517/2000 = 25.85\%$, whereas the separately generated mean seed-level conflict rate was reported as 26.19%. Those aggregates need not be arithmetically identical under distinct pooling, but no prompt/seed mapping supports that explanation. It remains unresolved; we do not invent a reconciliation or allocate empty outputs to cells.

Model-family labels are not nonexistent merely because source records are missing. Pythia, Qwen2.5 and Mistral are established families. However, "Mistral-Nemo-12B" is insufficient to identify a base versus instruction checkpoint, and none of these labels establishes a run. A fixed Python `hash()`-derived random generator can also vary across interpreter processes unless the hash seed is controlled; this is simulation reproducibility, not model reproducibility. Tight synthetic token budgets must not be described as measured training parity.

An additional staged extension contains halted jobs, not completed confirmation: its SFT baseline stops at 40/400 iterations without a final adapter, and its second PT seed has no completed metrics or validation output. These records are not counted as completed seeds. Format compliance does not establish scientific readiness.