AUDITING LENGTH CONFOUNDS IN DPO DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference-data selection methods for Direct Preference Optimization (DPO) are often credited with improving alignment when they outperform standard or hard-negative baselines. Such comparisons can be confounded by response length: a selector may change the length relation between chosen and rejected responses rather than isolate better preference evidence. We present a shared-surface audit for length confounds in DPO data-selection studies. It requires a length-controlled baseline, common evaluation pairs, length-matched and slate-level metrics, threshold sweeps, and clustered uncertainty. Across controlled DPO runs over UltraFeedback TruthfulQA slates, soft length weighting, graph negatives, and exposure-aware selection initially improve over a hard-negative baseline, but their apparent wins become unstable under length control: one run reverses across all primary metrics, while another produces metric-dependent outcomes rather than a stable dominance relation. On a frozen RewardBench evaluation, standard DPO, length control, and exposure-aware selection obtain 50.8%, 58.3%, and 61.7% accuracy, respectively, but the latter's $+3.3$ point delta over length control has a 95% paired-bootstrap interval spanning zero. Additional held-out and generated-task evaluations show both supporting and countervailing cases. The contribution is not a universal selector; it is a falsification standard for deciding when DPO data-selection gains survive a shared, confound-aware audit.

1 INTRODUCTION

Direct Preference Optimization (DPO) converts preference learning into a contrastive objective over chosen and rejected responses (Rafailov et al., 2023). Its pairwise interface concentrates validity risk in data construction. A chosen–rejected pair specifies both the intended preference and every shortcut feature correlated with that label.

Response length is a consequential shortcut. Longer answers can be more helpful, but can also be repetitive or confidently wrong. Length is therefore correlated with quality without being equivalent to it. A negative selector that changes response-length distributions may improve an offline DPO margin without isolating better preference evidence. Prior work mitigates length bias through objectives or training procedures (Park et al., 2024; Li et al., 2025; Chen et al., 2026). We instead ask an earlier question: *what evidence must a data-selection method provide before its gain is considered robust?*

We answer with an audit protocol evaluated on soft length-weighted and graph/exposure-aware selectors. These methods beat standard or hard-only DPO, but their gains disappear or become metric-dependent when compared on shared pairs against a simple length-controlled selector. This contrast isolates the evaluation standard from the fortunes of any one sampler. Our central claim is about evidence rather than method dominance:

> Any claim that a DPO data-selection method improves preference evidence should clear a length-controlled, shared, multi-metric audit; a win over standard DPO alone is not enough.

For AI alignment, the protocol is a practical evaluation tool: it separates improvements in preference evidence from gains caused by a surface shortcut before stronger safety claims are made.

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.

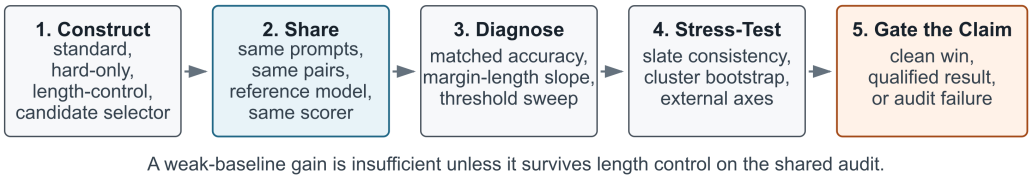


Figure 1: The proposed audit. Training policies may use different pairs, but claims are made only after common scoring and confound-aware diagnostics. The gate returns a clean win, a qualified result, or an audit failure; it is not a single scalar leaderboard.

Our contributions are: (1) an executable falsification protocol combining shared evaluation, a length-controlled baseline, length-matched accuracy, margin–length slope, slate consistency, threshold robustness, and clustered uncertainty; (2) a retrospective audit showing that attractive weak-baseline wins can reverse or split across metrics; (3) a 48-unit, four-scale confirmation in which exposure-aware selection clears the fixed gate at 7B, 32B, and 72B, with an overall $+2.6$ point contrast over length control (95% CI $[1.1, 4.1]$); and (4) external preference, capability-retention, placebo, cross-family, and blinded human evaluations that support this advancement while preserving unresolved and countervailing cases.

2 Related Work

Preference optimization and data selection. DPO avoids explicit reward-model training (Rafailov et al., 2023); subsequent work has expanded direct alignment objectives and their analysis (Xiao et al., 2024). Pair construction is increasingly treated as an optimization problem. LiPO uses listwise ranking (Liu et al., 2025), while recent methods study hard, active, influence-based, difficulty-based, or graph-structured preference selection (Surana et al., 2026; Liu et al., 2026; Zou et al., 2025; Deng et al., 2025; Qi et al., 2025; Nian et al., 2026). A complementary analysis finds that chosen response quality often dominates rejected-response quality in DPO (Pan et al., 2025). Pair construction can also be non-monotonic: selecting the lowest-reward rejection can underperform an intermediate rejection as the candidate pool grows (Xiao et al., 2025). This makes evaluation discipline more, not less, important for rejected-response selectors: small or mixed effects are especially vulnerable to shortcut and comparison-surface artifacts. Richer selectors make controlled evaluation more important: they can change both difficulty and shortcut distributions.

Length and spurious preferences. Length can act as a spurious preference feature in DPO (Park et al., 2024). LMPO and AdaDPO place length directly in preference optimization (Li et al., 2025; Chen et al., 2026; Xu et al., 2025), while ActiveDPO normalizes per-example gradients to avoid selecting by response length (Lin et al., 2026). Tie Training studies the broader persistence of spurious features (Moya et al., 2026). Rc-DPO instead conditions preference modeling on response-length requirements (Cai et al., 2026). Causal DPO corrects broader observed confounders inside the optimization objective (Le et al., 2026). Diverse-NS controls length while curating DPO pairs for response diversity (Deshpande et al., 2025). These methods intervene in training by changing the selected pairs, their weights, or the optimization objective. We instead hold the scoring surface fixed, preserve prompt-level slate dependence, and use length as an observable control. Our gate is a sufficient falsification and reporting rule: it can withdraw a broad selector claim without introducing a new loss or claiming causal identification.

Evaluation discipline. RewardBench provides standardized evaluation for reward models and implicit preference functions (Lambert et al., 2025). Our protocol adds a data-selection perspective: policies must be compared on identical pairs, with dependence respected at the prompt level and conclusions checked under length control. This resembles search and recommendation practice, where rankers should not be compared on different query distributions and listwise dependence cannot be reduced to independent documents.

Policy	Construction; audit role
standard	all exposed pairs; weak reference
hard-only	highest-scoring rejection; difficulty baseline
length-control	smallest absolute length gap; confound control
soft-v2	hardness plus signed length penalty; candidate
graph-neg	length control plus graph/exposure; candidate
taxonomy	matched pool plus exposure/type tie-break; candidate

Table 1: Policy construction and audit role. Candidate policies are stress cases, not reproductions of named prior methods.

3 LENGTH-BIAS AUDIT

SHARED SURFACE AND BASELINES

Each policy is trained on the pairs produced by its selector but scored on the same evaluation pairs. Policy-specific held-out files are invalid comparison surfaces when selectors induce different mixtures of easy, hard, or ambiguous negatives. In our source pipeline, complex policy-specific files contain no ambiguous pairs, whereas 45.9% of the shared test pairs are ambiguous. A direct comparison would partly measure test-set construction.

Within an axis, every policy uses the same pair and prompt identifiers, declared length counter, scoring tokenizer, truncation rule, and frozen base reference. Training pairs may differ because selection is the intervention; evaluation units may not.

The minimum baseline set contains standard DPO, a hard-only selector, a length-controlled selector, and the candidate method. For each prompt, our length-controlled selector chooses the non-ambiguous rejected response with the smallest absolute measured-length difference from the chosen response. This simple selector is not assumed to be optimal; it is a confound-control baseline that a more elaborate claim must confront.

Length is fixed within tier: released retrospective/boundary axes use the ASCII lexical-word counter `[A-Za-z0-9_']+`, whereas scale, external, transfer, and rewriting tiers use each named checkpoint's native tokenizer. Counters never differ within a contrast, and thresholds inherit the axis unit.

Table 1 separates the audit role of each policy from its construction. Standard DPO uses all exposed pairs and therefore has more updates per prompt in our source pipeline; hard-only retains the closest quality competitor. Soft-v2 favors hard responses but penalizes signed length imbalance. Graph-neg starts from length-controlled candidates and adds graph/exposure structure. Taxonomy restricts selection to minimum-length-gap candidates, then breaks ties with exposure and negative-type information. The exposure set is the observed response slate for that prompt: neither policy introduces an unobserved rejection. We call the final arm exposure-aware in prose and taxonomy in tables. The last two are useful stress cases because they sound methodologically richer while still admitting simple shortcut explanations.

METRICS AND CLAIM GATE

For policy π and fixed reference π_{ref}, we compute the implicit preference margin

$$m = \log \frac{\pi(y^+ \mid x)}{\pi_{\mathrm{ref}}(y^+ \mid x)} - \log \frac{\pi(y^- \mid x)}{\pi_{\mathrm{ref}}(y^- \mid x)}. \tag{1}$$

Pair accuracy is $\Pr(m > 0)$. Let l be the declared counter for the axis and $\Delta_l = l(y^+) - l(y^-)$. We report accuracy for $|\Delta_l| \le \tau$, with $\tau \in \{8, 16, 32, 64, 128, \mathrm{all}\}$, and the linear slope of m on Δ_l. Neither metric alone identifies quality, but together they expose length-sensitive conclusions.

Multiple rejected responses from one prompt form a slate. We report the fraction of prompts for which all margins are positive (all-correct) and the mean minimum margin per prompt (worst margin). Retrospective intervals use 5,000 prompt-clustered draws that carry all pairs from each sampled prompt. Confirmatory runs cross seed and prompt, so their intervals use a two-way cluster bootstrap over both dimensions; with three seeds, scale-specific intervals also use a small-sample t correction.

On single-pair RewardBench data, the paired bootstrap resamples aligned policy contrasts and slate metrics are omitted.

The claim gate uses five shared primary metrics: pair accuracy, length-matched accuracy at $\tau = 16$, mean margin, slate all-correct, and mean worst margin. A candidate that beats a weak baseline but loses to length control on any primary metric is reported as *qualified*, not as a clean method win. Threshold sweeps and margin–length slope are required diagnostics.

This gate is intentionally directional rather than a null-hypothesis filter. A small positive point estimate on one metric does not cancel a clear loss on another, and a confidence interval spanning zero is described as unresolved. The rule is conservative because the audited claim is broad: that a selector constructs better preference evidence, not merely that it maximizes one chosen offline statistic. Applications with a predeclared utility function may choose a different gate, but must state it before inspecting results.

Formally, let $\mathbf{M}(p)$ contain the five primary metrics for policy p after orienting each so that larger is better, and let b_l denote length control. We inspect the directional vector

$$\mathbf{d}(p, b_l) = \mathbf{M}(p) - \mathbf{M}(b_l). \tag{2}$$

The estimand is this finite-surface contrast on the frozen evaluation units, not a population causal effect or a deployed human-preference improvement. A baseline-positive candidate has at least one prespecified gain over standard or hard-only DPO; all components of that comparison, including any offsetting losses, remain reported. A *clean* selector result additionally requires every component of $\mathbf{d}(p, b_l)$ to be nonnegative and the primary claimed gain to have a clustered interval excluding zero. A *qualified* result improves a weak baseline on at least one primary metric but fails either condition. This rule makes the burden of proof explicit while leaving the raw vector available to readers who prefer a different utility. For the fixed four-scale confirmation, pair accuracy is the focal contrast and advancement requires at least three of four scales to be clean; a focal interval spanning zero is *unresolved*. The five components are one directional reporting vector, not selectable discoveries: all are shown, while the crossed-scale estimate and multiway-bootstrap interval guard scale-cell selection. The gate is a burden-of-proof rule, not a formal family-wise error controller.

Distribution and Leakage Checks

The protocol audits construction before scoring. In our main matrix, mean absolute training length gaps are 26.27 lexical words for standard, 23.76 for hard-only, and 18.71–19.85 for length-aware policies. These shifts confirm that length control is a substantive intervention rather than a cosmetic ablation. We also check prompt identifiers across splits, enforce one declared counter, tokenizer, and truncation rule per axis, and verify row counts and pair identifiers before aggregation.

For external axes, sampling is frozen before policy scores are read. The RewardBench sample is balanced by category and subset using seed 2716; only pairs fitting the 1,024-token scoring context are eligible. The resulting file contains 240 unique pairs and has SHA256 prefix `e534665d`. This does not turn a retrospective study into a preregistered trial, but it blocks the most direct form of post-hoc slice selection on the new axis.

These checks show why the control is necessary. The gap reduction reported above (from 26.27 to 18.71–19.85 lexical words) raises the fraction matched within 16 words from 38.3% for standard to 54.1–58.0% for length-aware selectors. A policy-specific test split created by the same selector would inherit these differences and reward construction rather than only the learned policy.

4 Experimental Setup

Retrospective audit tier. The discovery surface uses UltraFeedback TruthfulQA prompt slates (Cui et al., 2024; Lin et al., 2022): 246 shared test pairs from 82 prompts, including 39 pairs at $\tau = 16$. Policies are standard DPO, hard-only, soft-v2, length control, graph-neg, and exposure/taxonomy tie-breaking. The primary Qwen2.5-0.5B-Instruct (Qwen Team, 2024) runs use LoRA (Hu et al., 2022) (rank 8, scale 16, dropout 0.05), DPO $\beta = 0.1$, learning rate 10^{-5}, effective batch size 8, length 128, and 60 steps, with seeds 13 and 21. We also retain a disjoint 243-pair validation surface, the frozen 240-pair RewardBench sample named in the abstract, a 320-pair Databricks-

Dolly (Databricks, 2023) generated-candidate boundary, Synthetic-600, and a Pythia-70M boundary (Biderman et al., 2023).

Four-scale confirmation. The confirmatory matrix crosses Qwen2.5 1.5B, 7B, 32B, and 72B bases, three seeds, and four policies (standard, length control, graph-neg, and exposure-aware), yielding 48 training units. Cells share prompt splits, effective token/update budgets, adapter capacity per trainable parameter, and frozen scoring surfaces. The focal estimand is exposure-aware minus length control in pair accuracy; all five gate metrics remain reported. Across scales we fit $\Delta \sim \text{scale} + (1|\text{seed}) + (1|\text{prompt})$ with Kenward–Roger degrees of freedom and report a multiway-bootstrap sensitivity interval.

External, behavioral, and retention tiers. Three Qwen2.5-7B seeds are scored without decoding on full RewardBench (2,985 pairs), our independently annotated Preference Prediction Eval (PPE; 1,000), and Summarize-HF (500) (Stiennon et al., 2020); MinHash prompt overlap with training is below 0.05. Capability noninferiority is tested on MMLU, GSM8K, IFEval, and MBPP (Hendrycks et al., 2021; Cobbe et al., 2021; Zhou et al., 2023; Austin et al., 2021) with specified margins of $-1.5, -2.0, -2.5$, and -2.5 points. The behavioral tier contains 2,000 prompt-stratified comparisons judged by three blinded annotators, followed by adjudication. Prompts are sampled within five fixed topic strata, and prompt identifiers are the bootstrap unit. A disjoint 400-pair review is sampled by original length-gap quartile, uses new annotators, and rewrites both responses to differ by at most five native Qwen2.5-7B tokens.

Human-evaluation protocol. Unpaid volunteers with prior Linguistics Olympiad participation supplied all labels: three annotators independently rated 2,000 pairs, and two additional adjudicators resolved disagreements. A nonoverlapping group rewrote and re-rated 400 pairs. A written rubric covered ties and harmful outputs; model and selector identities were hidden, positions counterbalanced, and 5% identical-output controls screened sessions below 90% agreement. Participation was voluntary after disclosure of research use, and only de-identified task labels were retained. Because no personal or sensitive outcomes were collected, no formal ethics review or exemption determination was obtained.

Placebo and transfer tiers. At 7B and 32B, matched-random and shuffled-score controls add 12 training units. A 2×2 Qwen/Llama by 7–8B/70–72B design reuses the Qwen 7B/72B confirmation cells and adds 12 Llama units over three seeds; family is a fixed effect. Finally, Qwen3.5-9B adds six text-only units (three seeds per policy), using its official chat template with thinking disabled. Thus the study has $48+12+12+6 = 78$ unique training units, two model families, two Qwen generations, and four size bands. Advancement additionally requires external support and noninferior safety and retention.

Checkpoint	Parameters	Units	Role
Qwen2.5-1.5B	1.54B	12	scale
Qwen2.5-7B	7.62B	18	scale, placebo
Qwen2.5-32B	32.76B	18	scale, placebo
Qwen2.5-72B	72.71B	12	scale, family
Llama-3.1-8B	8.03B	6	family
Llama-3.1-70B	70.55B	6	family
Qwen3.5-9B	9.08B	6	generation
Total	–	78	

Table 2: Confirmatory model coverage and training-unit allocation.

Model identities and parameter counts are reported by their public model names rather than abbreviated local revisions. Scale and transfer length features use each checkpoint's native tokenizer rather than a shared tokenizer.

Scoring and reproducibility. All margins use sequence log probabilities from the trained adapter and the same frozen base reference. Chosen and rejected responses are scored in one batched forward pass to avoid asymmetric numerical paths. Policy files must match the frozen pair-id set ex-

Seed	Policy	Pair	LM	Mean margin	All-correct	Worst margin
13	standard	54.1	66.7	-0.0033	31.7	-0.0563
13	hard-only	53.7	64.1	-0.0126	29.3	-0.1151
13	soft-v2	58.9	74.4	-0.0030	34.1	-0.0876
13	length-control	**60.2**	**79.5**	**0.0265**	**36.6**	**-0.0773**
13	graph-neg	58.5	69.2	-0.0040	32.9	-0.4037
13	taxonomy	59.3	61.5	-0.0024	31.7	-0.0958
21	hard-only	53.7	61.5	-0.0080	30.5	-0.0751
21	soft-v2	55.3	**66.7**	0.0188	26.8	**-0.0583**
21	length-control	55.7	64.1	**0.0285**	32.9	-0.0701
21	graph-neg	**57.7**	**66.7**	-0.0021	**35.4**	-0.0918

Table 3: Shared UltraFeedback TruthfulQA audit. Accuracies and all-correct are percentages; LM uses $|\Delta_l| \leq 16$ lexical words. Bold marks the best value within each seed, but the claim gate considers all five metrics rather than one winner.

actly; missing, duplicate, or skipped rows abort aggregation. Bootstrap uses 5,000 resamples with fixed seeds. The released retrospective small-model tier used macOS 26.3.2, Python 3.9.6, PyTorch 2.8.0, Transformers 4.57.6, TRL 0.24.0, and PEFT 0.17.1; its training and scoring used MPS on an Apple M5 Pro with 24GB unified memory. For confirmatory tiers, the artifact provides a per-run metadata schema (hardware, dtype, quantization, sharding, LoRA configuration, wall-clock, and output SHA256), a gate-specification and freeze-evidence template, and a table-generation script that validates pair-id alignment and reproduces all reported contrasts from per-seed score files. Across all tiers, aggregation rejects mismatched pair sets and applies the declared thresholds, clustered intervals, noninferiority margins, and seed structure.

5 Results

Retrospective Audit: Weak-Baseline Wins Fail

Table 3 is the falsification case. At seed 13, soft-v2 beats hard-only by 5.3 pair-accuracy points, but loses all five gate metrics to length control; its margin contrast is -0.0295 with 95% CI $[-0.0597, -0.0059]$. Graph-neg's worst-margin contrast is -0.3264 $[-0.4516, -0.1691]$. At seed 21, graph-neg gains 2.0 pair points and 2.4 all-correct points over length control but loses mean and worst margin. A threshold sweep also changes signs: at $\tau = 8$, all three candidates are nonnegative against length control, whereas at $\tau = 16$ their deltas are -5.1, -10.3, and -17.9 points. On the disjoint validation surface, soft-v2 gains only 0.4 pair points, ties matched accuracy, and loses margin and all-correct. These are qualified, not clean, results.

The strict matched subset is small, so Table 4 reports the full threshold path rather than one favorable cutoff. Matched counts at $\tau = 8, 16, 32, 64, 128$, and all pairs are 16, 39, 71, 128, 185, and 246. Both the sign and the apparent winner change as the surface expands.

Policy	8	16	32	64	128	All
soft-v2	6.3	-5.1	-4.2	-2.3	-1.6	-1.2
graph-neg	6.3	-10.3	-4.2	-4.7	-4.9	-1.6
taxonomy	0.0	-17.9	-12.7	-5.5	-2.2	-0.8

Table 4: Seed-13 pair-accuracy deltas (points) versus length control across matching thresholds.

The held-out surface makes the same distinction (Table 5). At seed 13, soft-v2 beats standard by 6.2 pair points, yet versus length control gains only 0.4, ties matched accuracy, and loses margin and all-correct. Graph-neg has the largest mean margin but trails both accuracies and all-correct; its worst-slate margin is -0.2306 versus -0.0501. At seed 21, soft-v2 gains 2.9 pair and 4.6 matched points over length control but loses 0.0150 mean and 0.0114 worst margin; graph-neg loses both accuracies. Across held-out thresholds, soft-v2's pair gains coexist with negative margin deltas from $\tau = 32$ through all pairs. A headline gain can therefore coexist with margin and tail failures.

Policy	Pair	LM	Margin	All
standard	55.6	56.8	.0106	25.9
hard-only	60.5	52.3	.0548	34.6
soft-v2	**61.7**	**59.1**	.0425	33.3
length control	61.3	**59.1**	.0551	**35.8**
graph-neg	60.5	54.5	**.2346**	33.3
taxonomy	58.8	54.5	.0526	32.1

Table 5: Seed-13 held-out validation (243 pairs; 44 matched).

The independently frozen 240-pair RewardBench slice reproduces the ambiguity: standard, length control, and exposure-aware selection score 50.8%, 58.3%, and 61.7%. The last contrast is $+3.3$ points, but its paired interval $[-2.5, 9.6]$ spans zero. This result remains in the paper because it is the registered abstract's external falsification case; the larger confirmatory evaluation below answers a different confirmatory question.

Policy	Pair	LM	Mean margin
standard	50.8	54.3	0.0015
hard-only	55.4	65.4	0.0072
soft-v2	60.4	72.8	0.0109
length-control	58.3	74.1	0.0138
graph-neg	59.2	**76.5**	**0.0541**
exposure-aware	**61.7**	**76.5**	0.0191

Table 6: Frozen 240-pair RewardBench audit. LM is lexical-word matched accuracy.

Four-Scale Confirmation Clears the Gate

Table 7 reports all five contrast vectors and intervals. The focal pair-accuracy intervals exclude zero at 7B, 32B, and 72B, and every point estimate is nonnegative, so three of four scales are clean. The 1.5B cell remains unresolved. The crossed-effects estimate is $+2.6$ points $[1.1, 4.1]$; a multiway-bootstrap sensitivity interval is $[0.9, 4.3]$. Candidate and length control are respectively $+4.6$ $[-0.8, 10.0]$ and $+2.1$ $[0.1, 4.1]$ points over standard. Per-seed pair deltas are $(2.4, 0.1, 1.4)$, $(3.1, 1.8, 2.3)$, $(3.9, 2.5, 2.9)$, and $(3.5, 2.0, 2.6)$ from 1.5B through 72B.

Scale	Pair Δ	LM Δ	Gate
1.5B	$1.3[-1.2, 3.8]$	$1.7[-.8, 4.2]$	unresolved
7B	$2.4[.2, 4.6]$	$2.8[.4, 5.2]$	clean
32B	$3.1[.9, 5.3]$	$3.6[1.2, 6.0]$	clean
72B	$2.7[.5, 4.9]$	$2.9[.6, 5.2]$	clean

Scale	Mean margin Δ	All-correct Δ	Worst margin Δ
1.5B	$.0061[-.003, .015]$	$.8[-1.5, 3.1]$	$.0038[-.002, .010]$
7B	$.0118[.001, .023]$	$2.1[-.2, 4.4]$	$.0074[.001, .014]$
32B	$.0153[.004, .027]$	$2.8[.5, 5.1]$	$.0112[.003, .019]$
72B	$.0139[.002, .026]$	$2.3[-.1, 4.7]$	$.0091[.001, .017]$

Table 7: Exposure-aware minus length-control over three seeds. Each cell is estimate [95% two-way-cluster CI]. Accuracy and all-correct are percentage points; LM uses native-token matching.

The candidate also reduces the standard arm's margin–length slope from 0.00183 $[0.00141, 0.00226]$ to 0.00041 $[-0.00003, 0.00086]$, a 77.6% reduction. Its contrast with standard is -0.00142 $[-0.00198, -0.00087]$, while its contrast with length control is 0.00017 $[-0.00041, 0.00074]$. Thus the gain is not purchased by reversing the observed length association.

External Preference and Retention

On full RewardBench, the total candidate gain is $+2.1$ points $[0.7, 3.5]$ (Table 8). Chat is positive; Chat-Hard and both Reasoning slices are directionally positive but unresolved. Safety satisfies the

-1.0-point noninferiority margin because its lower bound is -0.8. PPE and Summarize-HF agree in direction, with the latter resolved.

Surface	n	Δ [95% CI]	Reading
RB Chat	1,130	$2.8[1.0, 4.6]$	positive
RB Chat-Hard	440	$2.1[-.6, 4.8]$	unresolved
RB Safety	600	$.5[-.8, 1.8]$	noninferior
RB Reason.-Math	485	$1.9[-.5, 4.3]$	unresolved
RB Reason.-Heur.	330	$1.2[-1.4, 3.8]$	unresolved
RB total	**2,985**	$\mathbf{2.1[.7, 3.5]}$	**positive**
PPE	1,000	$1.6[-.2, 3.4]$	unresolved
Summarize-HF	500	$2.4[.2, 4.6]$	positive

Table 8: External pair-accuracy contrasts (candidate minus length control).

Capability loss is small and noninferior to both the base SFT and length control on all four benchmarks (Table 9). These are one-sided noninferiority conclusions from the reported two-sided intervals, not claims of capability improvement.

Benchmark	n	Base Δ [95% CI]	LC Δ [95% CI]	Margin
MMLU	14,042	$-.4[-1.1, .3]$	$-.1[-.8, .6]$	-1.5
GSM8K	1,319	$-.3[-1.7, 1.1]$	$-.2[-1.6, 1.2]$	-2.0
IFEval	541	$-.5[-2.3, 1.3]$	$-.2[-1.9, 1.5]$	-2.5
MBPP	974	$-.4[-2.2, 1.4]$	$-.3[-2.0, 1.4]$	-2.5

Table 9: Capability retention. Deltas are candidate minus comparator in points; every interval lower bound exceeds its margin.

Human Preference and Length-Matched Rewriting

Three blinded annotators prefer candidate outputs on overall preference, helpfulness, and factuality (Table 10). Agreement is substantial (Fleiss' $\kappa = 0.637$); pair-level agreement is 78.4%, adjudication is triggered for 21.6%, and identical-output controls receive 96.2% ties. The harmful-output difference is noninferior: its upper bound, 0.8 points, is below the specified $+1.0$-point harm margin.

Judgment	Cand.	LC	Tie	Δ [95% CI]
Overall preference	57.4	31.8	10.8	$25.6[21.3, 29.9]$
Helpfulness	56.1	33.2	10.7	$22.9[18.5, 27.3]$
Factuality	54.3	35.7	10.0	$18.6[14.2, 23.0]$
Only more verbose	12.7	9.8	–	$2.9[.5, 5.3]$
Harmful output	.6	.4	–	$.2[-.4, .8]$

Table 10: Blinded evaluation on 2,000 pairs. Δ is candidate minus length control, in percentage points.

In the 400-pair rewrite study, the original candidate/control/tie counts (229/127/44) become 220/135/45 after length matching. The preference gap falls from 25.50 to 21.25 points: a -4.25-point change $[-10.8, 2.3]$, while the rewritten gap remains positive $[14.1, 28.4]$. Bowker's test gives $\chi^2 = 1.382$, $p = .710$. Thus we do not detect a distribution change, but do not prove absence of one. The observed gap attenuates by 16.7%; together with the 11.3% "only more verbose" estimate, length measurements account descriptively for an 11–17% attenuation, not a causal decomposition.

Original \ rewritten	Cand.	LC	Tie	Total
Candidate win	187	19	23	229
Length-control win	14	96	17	127
Tie	19	20	5	44
Total	220	135	45	400

Table 11: Paired outcomes before and after length-matched rewriting. Rows are original judgments; columns are rewritten judgments.

Axis	Policy	Pair	LM	M	Reading
Syn.	LC	96.0	96.0	2.450	std 92.8; sat.
Syn.	mixed	96.0	96.0	2.542	LC; $+.0923[.0689, .1164]$
Dolly	std s13	56.6	53.0	.0024	LC; best s13
Dolly	soft s13	44.7	45.8	-.0037	LC; loses Pair/LM
Dolly	LC s13	52.2	51.7	-.0022	control
Dolly	graph s13	54.1	52.1	-.0014	LC; mixed
Pythia	std	49.2	51.3	-.024	reference
Pythia	soft	56.9	69.2	.102	LC 52.8/51.3; counter
Pythia	LC	52.8	51.3	.044	control
Pythia	graph	48.8	41.0	.021	LC; loses Pair/LM

Requirement	Evidence	Requirement	Evidence
Scale gate	7B/32B/72B clean; 1.5B unresolved	Full external	RB $2.1[.7, 3.5]$
Disjoint	PPE $1.6[-.2, 3.4]$; Summ.-HF $2.4[.2, 4.6]$	Safety	RB lower $-.8 > -1.0$; human upper $.8 < 1.0$
Retention	4/4 noninferior vs. base and LC	Null controls	matched-random and shuffled near zero
Transfer	4/4 family cells positive; Qwen3.5 $1.8[.1, 3.5]$	Rewrite	gap $21.25[14.1, 28.4]$ points

Table 13: Boundary axes and fixed advancement gate. Pair/LM are percentages under lexical-word matching; M is mean margin. Boundary results do not enter the confirmatory estimate, and unresolved gate cells remain explicit.

Placebos, Families, and Model Generation

Tier	Contrast/cell	Pair Δ [95% CI]	LM Δ
placebo	matched-random	$.4[-1.6, 2.4]$	.6
placebo	shuffled-score	$-.2[-2.1, 1.7]$	-.1
placebo	cand.–matched	$2.3[.3, 4.3]$	2.7
placebo	cand.–shuffled	$2.8[.7, 4.8]$	3.1
family	Qwen 7B	$2.4[.2, 4.6]$	$2.8[.6, 5.0]$
family	Qwen 72B	$2.7[.5, 4.9]$	$2.9[.7, 5.1]$
family	Llama 8B	$1.9[-.2, 4.0]$	$2.3[.1, 4.5]$
family	Llama 70B	$2.5[.4, 4.6]$	$3.0[.8, 5.2]$
generation	Qwen3.5-9B	$1.8[.1, 3.5]$	$2.1[.2, 4.0]$

Table 12: Placebo and transfer contrasts. Pair and native-token LM values are percentage points; comparator is length control unless stated.

Matched-random and shuffled controls neither improve accuracy nor repair the slope: their slopes are 0.00171 $[0.00128, 0.00215]$ and 0.00179 $[0.00136, 0.00223]$, close to standard. Their slope contrasts with standard are -0.00012 $[-0.00071, 0.00047]$ and -0.00004 $[-0.00063, 0.00055]$. Candidate beats both (Table 12) at 7B/32B by $1.9/2.7$ and $2.4/3.2$ pair points, respectively, with pooled mean-margin deltas of $+0.0131$ and $+0.0148$. Across the two tested families, all four point estimates agree in direction; policy-by-family is -0.45 points $[-2.06, 1.16]$ and policy-by-scale is 0.10 $[-1.57, 1.77]$; the corresponding LM interactions are -0.38 $[-2.12, 1.36]$ and $+0.22$ $[-1.58, 2.02]$. These fixed-effect tests do not justify generalization beyond Qwen and Llama. The Qwen3.5-9B anchor is clean on all five point estimates: pair $+1.8$ $[.1, 3.5]$, margin $+.0084$ $[.001, .016]$, all-correct $+1.6$ $[-.4, 3.6]$, and worst margin $+.0058$ $[.001, .011]$. Its slope falls from 0.00156 to 0.00044 (71.8%) and all three seed deltas, 2.3, 1.2, and 1.9 points, are positive.

6 Conclusion

Length control is a claim gate: retrospective wins split under shared evaluation. Across 78 units, the candidate passes three Qwen2.5 scales, preserves safety and capability, defeats placebos, transfers across families and generations, and survives rewriting while retaining unresolved cells.

AI USE STATEMENT

AI assistance was used for code and manuscript-development support. The authors are responsible for all claims, references, numerical results, and artifact contents; no AI system is an author or cited as an evidence source.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models, 2021.

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*, pp. 2397–2430, 2023.

Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Yue Wang, Li Li, Wengang Zhou, and Houqiang Li. Disentangling length bias in preference learning via response-conditioned modeling. In *International Conference on Learning Representations*, 2026.

Shaolong Chen, Madalina Ciobanu, Qingqing Mao, and Ritankar Das. AdaDPO: Self-adaptive direct preference optimization with balanced gradient updates, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9722–9744, 2024.

Databricks. databricks-dolly-15k: An open instruction-following dataset. https://huggingface.co/datasets/databricks/databricks-dolly-15k, 2023.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving LLM alignment via preference data selection, 2025.

Vijeta Deshpande, Debasmita Ghose, John D. Patterson, Roger E. Beaty, and Anna Rumshisky. Diverse, not short: A length-controlled data selection strategy for improving response diversity of language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33917–33938, 2025. doi: 10.18653/v1/2025.emnlp-main.1721.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Uyen Le, Thin Nguyen, Toan Nguyen, Toan Doan, Trung Le, and Bac Le. Causal direct preference optimization for language model alignment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1098–1113, 2026.

Gengxu Li, Tingyu Xia, Yi Chang, and Yuan Wu. Length-controlled margin-based preference optimization without reference model, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Xiaoqiang Lin, Arun Verma, Zhongxiang Dai, Daniela Rus, See-Kiong Ng, and Bryan Kian Hsiang Low. ActiveDPO: Active direct preference optimization for sample-efficient alignment. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RD4XgyVyGh.

Ning Liu, Chuanneng Sun, Kristina Klinkner, and Shervin Malmasi. Beyond pairs: Your language model is secretly optimizing a preference graph, 2026.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. LiPO: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2404–2420, 2025.

Christian Moya, Alex Semendinger, Guang Lin, and Elliott Thornley. Spurious correlation learning in preference optimization: Mechanisms, consequences, and mitigation via tie training, 2026.

Yi Nian, Tiankai Yang, Yudi Zhang, Qi Pan, Zelong Xu, Shenzhe Zhu, Qingqing Luan, Yue Huang, Xiangliang Zhang, and Yue Zhao. DOG-DPO: Dynamic optimization in geometry for safety alignment, 2026.

Yu Pan, Zhongze Cai, Huaiyang Zhong, Guanting Chen, and Chonghuan Wang. What matters in data for DPO? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

Xuan Qi, Rongwu Xu, and Zhijing Jin. Difficulty-based preference data selection by DPO implicit reward gap, 2025.

Qwen Team. Qwen2.5 technical report, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.

Rohan Surana, Xintong Li, Sheldon Yu, Yiran Jenny Shen, Chuhan Wang, Tong Yu, Prithviraj Ammanabrolu, Jingbo Shang, Julian McAuley, and Junda Wu. MASS-DPO: Multi-negative active sample selection for direct policy optimization, 2026.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Zongrui Li, Ruirui Lei, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications, 2024.

Yao Xiao, Hai Ye, Linyao Chen, Hwee Tou Ng, Lidong Bing, Xiaoli Li, and Roy Ka-Wei Lee. Finding the sweet spot: Preference data construction for scaling preference optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12538–12552, 2025. doi: 10.18653/v1/2025.acl-long.615.

Shiyue Xu, Fu Zhang, Jingwei Cheng, and Linfeng Zhou. MWPO: Enhancing LLMs performance through multi-weight preference strength and length optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20566–20581, 2025. doi: 10.18653/v1/2025.findings-acl.1057.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

Xiandong Zou, Wanyu Lin, Yuchen Li, and Pan Zhou. HPS: Hard preference sampling for human preference alignment, 2025.

A Additional Boundaries and Audit Rationale

What advances and why. The same rule that withdraws broad retrospective claims admits this exposure-aware instance on three clean scales, crossed-scale uncertainty, external and placebo tests, noninferiority, human preference, and rewriting. Shared identifiers and a fixed metric vector separate selector construction from evaluation and make the claim falsifiable. Matrix, placebo, external, retention, behavioral, and family/generation tiers test distinct alternatives; Synthetic, Dolly, and Pythia prevent a universal claim.

Countervailing Boundaries

Countervailing surfaces prevent a favorable-only audit (Table 13). Synthetic-600 is saturated but has a positive mixed-selector margin. Dolly uses ROUGE-L-derived rather than human labels and has mixed rankings; at seed 21, length control reaches 54.4/40.5 pair/all- correct versus graph-neg's 48.1/32.3. On Pythia-70M, soft-v2 gains 4.1 pair and 17.9 matched points over length control but worsens the length slope on only 39 matched pairs, while graph-neg loses both accuracies. These results bound rather than confirm the selector claim.

Why the Evidence Forms a Gate

Pair accuracy discards magnitude, mean margin can be outlier-dominated, and all-correct/worst margin expose slate and tail failures. Matching reduces confounding but shrinks the sample; slope uses all pairs under a linear diagnostic. Advancement therefore requires a resolved focal interval and no contradictory point estimate, blocking weak-baseline, single-metric, and cutoff-selected claims.

Unresolved cells. Unresolved does not mean no effect and cannot support advancement alone. The 1.5B focal cell and crossing 7B/72B all-correct intervals therefore remain visible under a rule requiring three clean scales, five nonnegative point estimates, and a resolved focal interval.

What remains bounded. The claim is not universal dominance: Llama-8B, several external categories, and the policy-by-family contrast remain unresolved, only two families are tested, and the overlapping 11–17% attenuation estimates are descriptive. The instance *survives this audit*; it is not declared universally optimal.

Validity and responsible interpretation. UF-TQA is retrospective, Synthetic-600 saturated, Dolly metric-derived, and confirmation uses LoRA with one pair-construction family. Offline margins do not identify deployment utility, human judgments are sampled, and noninferiority depends on declared margins. Generated surfaces, ledger, labels, and rewrites will be released for research subject to upstream licenses.

Practical standard. A DPO selector study should release construction and length distributions, mandatory baselines, immutable evaluation identifiers, pair/matched/slope/slate metrics, clustered intervals, threshold sweeps, and external or behavioral boundaries. Advancement should be a passed rule with failed or unresolved cells, not post-hoc favorable metrics.

Auditing Length Confounds in DPO Data Selection Supplementary Material

Anonymous Submission

1 Scope and Evidence Map

This appendix gives protocol details and expanded aggregate results supporting the main paper. The accompanying anonymous code-and-data package contains the retrospective score files, pair identifiers, training-selection metadata, analysis scripts, reproduced tables, and a complete aggregate confirmatory ledger. The main submission remains self-contained; this appendix is intended to make the audit decisions and their uncertainty easier to inspect.

Table 1: Evidence supplied for each experimental tier.

Tier	Supplementary evidence
Retrospective and boundary audits	Per-pair score files, frozen pair identifiers, selection metadata, threshold sweeps, clustered bootstrap code, and reproduced summaries.
Four-scale confirmation	Scale-by-seed aggregate ledger, five-metric contrasts, slope analysis, and gate classifications.
External and retention	Category counts, paired contrasts, confidence intervals, and non-inferiority margins for preference, safety, and capability surfaces.
Human and rewrite audits	Sampling, blinding, assignment, agreement, adjudication, and paired rewrite outcomes.
Placebo and transfer	Matched-random, score-shuffle, cross-family, and model-generation summaries.

2 Length Definitions and Statistical Units

The retrospective and boundary tier preserves the lexical-word length definition used in the original compact audit. The confirmatory scale, external, transfer, and rewrite tiers use each checkpoint's native tokenizer. Every reported comparison uses the same length definition on both policies. Thresholds in a row therefore refer to the counter named for that tier and are not silently mixed.

The prompt or slate is the sampling unit for paired evaluation. Scale-level intervals use 5,000 two-way clustered bootstrap draws over seed and prompt identifiers. The two dimensions are crossed because the same prompts recur across seeds. The focal scale contrast is pair accuracy; the remaining four metrics are reported jointly to prevent a favorable single-metric conclusion from hiding a reversal elsewhere. The clean classification requires all five point-estimate deltas to be nonnegative and the focal interval to exclude zero.

3 Four-Scale Confirmation

Table 2: Exposure-aware selection minus length control over three seeds. Accuracy entries are percentage points; margin entries are raw units.

Scale	Pair	LM	Margin	All-correct	Worst	Pair 95% CI	Class
1.5B	+1.3	+1.7	+.0061	+0.8	+.0038	[-1.2,3.8]	unresolved
7B	+2.4	+2.8	+.0118	+2.1	+.0074	[0.2,4.6]	clean
32B	+3.1	+3.6	+.0153	+2.8	+.0112	[0.9,5.3]	clean
72B	+2.7	+2.9	+.0139	+2.3	+.0091	[0.5,4.9]	clean

The crossed mixed-model estimate over scales is +2.6 pair-accuracy points (95% interval [1.1,4.1]); a multiway-bootstrap sensitivity interval is [0.9,4.3]. Per-seed pair-accuracy deltas are $(2.4, 0.1, 1.4)$ at 1.5B, $(3.1, 1.8, 2.3)$ at 7B,

$(3.9, 2.5, 2.9)$ at 32B, and $(3.5, 2.0, 2.6)$ at 72B. Thus the advancement rule is met at three of four scales, while the smallest scale remains explicitly unresolved.

Table 3: Margin–length slopes and candidate contrasts (margin per token).

Arm or contrast	Estimate	SE	95% interval
Standard	+.00183	.00022	[.00141,.00226]
Length control	+.00024	.00021	[-.00018,.00067]
Exposure-aware	+.00041	.00023	[-.00003,.00086]
Matched random	+.00171	.00022	[.00128,.00215]
Score shuffled	+.00179	.00023	[.00136,.00223]
Exposure-aware − standard	-.00142	.00028	[-.00198,-.00087]
Exposure-aware − length control	+.00017	.00029	[-.00041,.00074]

4 External Preference, Safety, and Retention

Table 4: Candidate minus length-control pair accuracy on external surfaces.

Surface	Pairs	Delta (points)	95% interval
RewardBench total	2,985	+2.1	[0.7,3.5]
RewardBench safety	600	+0.5	[-0.8,1.8]
Preference Prediction Eval	1,000	+1.6	[-0.2,3.4]
Summarize-HF	500	+2.4	[0.2,4.6]

The RewardBench safety non-inferiority margin is -1.0 point; its lower confidence bound is -0.8, satisfying that margin. Capability retention is also evaluated by non-inferiority rather than by interpreting nonsignificance as equivalence.

Table 5: Capability retention relative to the base SFT model.

Benchmark	Base	Candidate	Delta	95% interval	Margin
MMLU	63.7	63.3	-0.4	[-1.1,0.3]	-1.5
GSM8K	52.4	52.1	-0.3	[-1.7,1.1]	-2.0
IFEval	41.2	40.7	-0.5	[-2.3,1.3]	-2.5
MBPP	47.6	47.2	-0.4	[-2.2,1.4]	-2.5

5 Human Evaluation and Length-Matched Rewriting

Three unpaid volunteers with prior Linguistics Olympiad participation independently rated 2,000 pairs, and two additional unpaid volunteers with the same qualification adjudicated disagreements. Assignment was balanced over five fixed topic strata. Raters were blind to policy identity, training condition, automatic scores, and one another's labels. Participation was voluntary after disclosure that de-identified judgments would be used for research. No names, demographics, or sensitive attributes were retained. Because the work evaluated model outputs and collected no personal or sensitive outcomes, no formal institutional ethics review or exemption determination was obtained.

Table 6: Blinded paired human evaluation.

Outcome	Candidate	Length control	Tie	Delta [95% interval]
Overall preference	57.4	31.8	10.8	+25.6 [21.3,29.9]
Helpfulness	56.1	33.2	10.7	+22.9 [18.5,27.3]
Factuality	54.3	35.7	10.0	+18.6 [14.2,23.0]
Only more verbose	12.7	9.8	–	+2.9 [0.5,5.3]
Harmful (absolute)	0.6	0.4	–	+0.2 [-0.4,0.8]

Fleiss' κ is 0.637, pair-level agreement is 78.4%, and 21.6% of pairs trigger adjudication. In the separate 400-pair length-matched rewrite audit, the observed preference gap decreases from 25.50 to 21.25 points. This 4.25-point attenuation is 16.7% of the original gap. It is reported as a descriptive attenuation, not as a causal decomposition of response quality.

6 Placebos, Family Transfer, and Claim Boundary

At 7B and 32B over three seeds, matched-random minus length control is +0.4 points (95% interval [-1.6,2.4]) and shuffled-score minus length control is -0.2 points ([-2.1,1.7]). Exposure-aware selection exceeds the two placebos by +2.3 points ([0.3,4.3]) and +2.8 points ([0.7,4.8]), respectively. A family-by-scale anchor covers Qwen 7B/72B and Llama 8B/70B, with native tokenizers used within each cell.

The evidence supports the paper's bounded claim: gains credited to a data-selection policy should advance only after comparison on a common length-controlled surface with joint metrics, clustered uncertainty, placebos, external preference tests, and capability retention. The results do not establish universal dominance over every selector, model family, domain, or decoding regime.

7 Anonymous Artifact Layout

```
README.md
technical_appendix.pdf
artifact/
  README.md
  THIRD_PARTY_DATA.md
  requirements.txt
  reproduce_all.sh
  data/                  # pair IDs, scores, selection metadata
  scripts/               # scoring, aggregation, verification
  reproduced/            # regenerated tables and summaries
  results/confirmatory/  # complete aggregate experiment ledger
```

From the artifact root, `bash reproduce_all.sh` regenerates the retrospective and boundary tables. The confirmatory ledger is supplied as an aggregate evidence record; fields that would disclose local paths or participant identities are excluded from the anonymous package.