Auditing Option-Position Effects in Multiple-Choice Evaluation of Domain-Fine-Tuned Language Models

Anonymous ACL submission

Abstract

Multiple-choice question (MCQ) accuracy is routinely used to compare domain-specific supervised fine-tuning (SFT) configurations, but such comparisons assume that a configuration's apparent gain does not depend on the correct option's position. We audit this assumption and find that domain SFT can shift a model's position preferences without changing its permutation-debiased accuracy, and that this divergence is large enough to alter method comparisons. A paired MedMCQA contrast (same items and hyperparameters, different supervision formats) yields two Qwen3-4B variants that share 51.8% permutation-debiased accuracy yet diverge in their predicted-position distributions. The MCQ-supervised variant shifts its most-predicted position share by +18.3 percentage points (pp) from the base checkpoint; the Open-QA variant by +3.5 pp. In a 40-run Qwen3-4B campaign on 3GPP telecommunications data, supplemented by cross-architecture checks on Mistral-7B-Instruct and LLaMA-3.1-8B-Instruct, we find that 39% of the raw-accuracy change after SFT (95% CI: 33–46%) is sensitive to option order. On the homogeneous 3GPP pool, the IFD-over-Random gap, largest among five selectors under default reporting, falls from +4.43 pp under raw, mismatched-budget scoring to +1.64 pp under matched-budget permutation-debiased scoring, no longer surviving the pre-specified correction. We recommend reporting raw accuracy, permutation-debiased accuracy, and predicted-position distributions together as a position-aware protocol for fixed checkpoints, not a model intervention. We release 3GPP-Eval V3.2, 11,846 QA pairs, and reporting code.

1 Introduction

Multiple-choice question (MCQ) benchmarks are routinely used to compare domain-specific supervised fine-tuning (SFT) configurations. In telecommunications, for example, LLM studies report raw MCQ accuracy as a primary outcome of domain adaptation (Maatouk et al., 2024; Zou et al., 2025; Kumar et al., 2025). Because these evaluations usually keep option order fixed across checkpoints, a raw-accuracy gain is read as evidence that a configuration is better adapted to the domain.

That reading relies on a measurement assumption: conclusions about a fine-tuning configuration should not depend on where the correct option appears. Under a fixed option order, however, accuracy can reflect both content-based evidence and a checkpoint's tendency to select particular positions. If reordering options changes which checkpoint appears better, the apparent gain may reflect position preference as well as configuration-level improvement.

Prior work makes this concern plausible. LLMs can favor particular option positions during inference (Zheng et al., 2024; Zhou et al., 2024), and fine-tuning can amplify that tendency, increasing first-option preference on intent classification with 54–150 options (Raimondi and Gabbrielli, 2025). However, it has not been tested whether the shift introduced by domain SFT is large enough to alter configuration comparisons on standard four-option MCQ benchmarks.

In this paper, we audit the measurement assumption in three steps. First, a paired MedMCQA contrast asks whether predicted-position behavior—how a checkpoint distributes its predictions across option positions—can diverge even when permutation-debiased accuracy is the same. Second, a 40-run Qwen3-4B campaign on 3GPP, with cross-architecture checks on Mistral-7B-Instruct and LLaMA-3.1-8B-Instruct, measures the scale of position-sensitive score change in the target domain. Third, we compare five data-selection methods under matched-budget, permutation-debiased scoring, testing whether position-aware reporting changes support for a

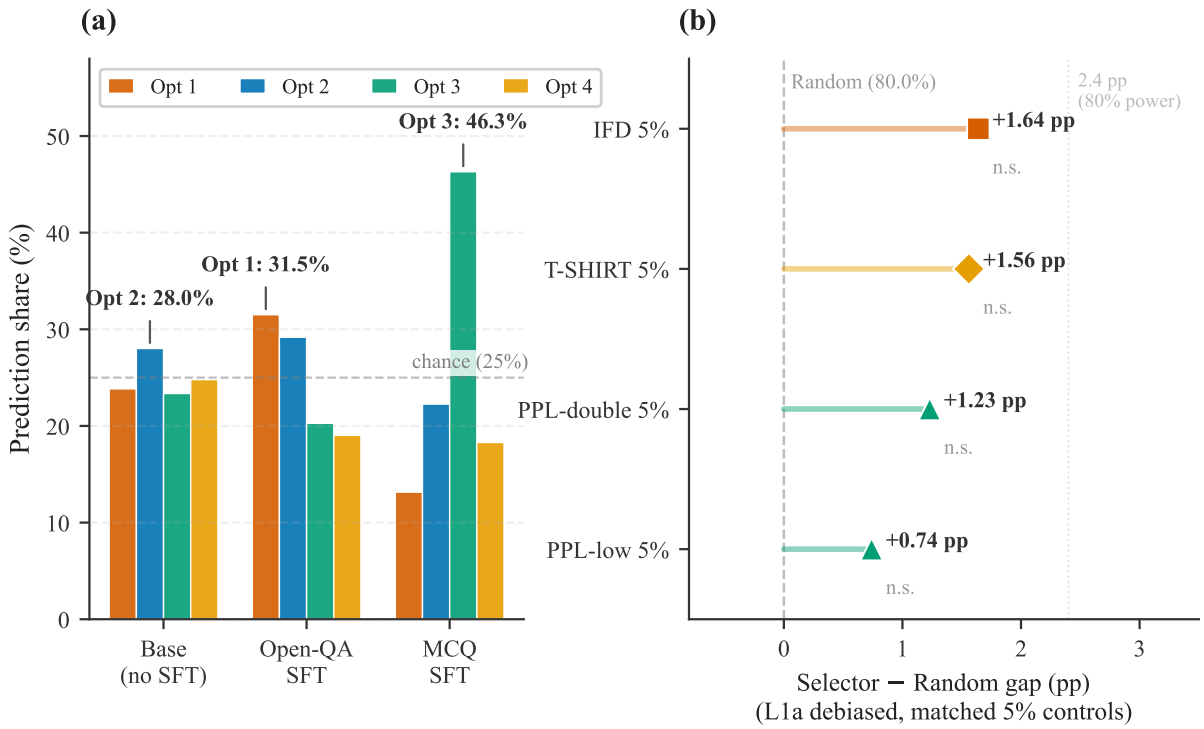

Figure 1: **Overview: identical debiased accuracy with divergent position behavior (a), and small, non-significant selector gaps under matched controls (b). (a)** MedMCQA paired contrast: MCQ and Open-QA variants of Qwen3-4B both reach 51.8% permutation-debiased accuracy, but shift MaxPred% (most-predicted position share) by +18.3 pp and +3.5 pp from the base, respectively. **(b)** Matched-budget, permutation-debiased selector–Random gaps on 3GPP L1a for the four plotted 5% runs. Gaps range from +0.7 to +1.6 pp; none is Bonferroni-significant (k=12). Dashed line: 80%-power threshold (2.4 pp).

method-level conclusion.

Figure 1 summarizes the central pattern. The paired contrast yields two variants with the same permutation-debiased accuracy but sharply different predicted-position behavior. On the 3GPP-Eval benchmark, 39% of the raw accuracy change after SFT (95% CI: 33–46%) does not persist after averaging over option order. This option-order instability matters for method comparison in the homogeneous 3GPP pool: the gap between the IFD data-selection method (Li et al., 2024b) and the Random baseline falls from +4.43 pp under raw, mismatched-budget reporting to +1.64 pp under matched-budget permutation-debiased scoring, no longer surviving the pre-specified multiple-comparison correction.

This paper makes three contributions. First, we operationalize position stability as a testable assumption, with permutation-debiased accuracy and predicted-position diagnostics as paired indicators. Second, we show that permutation-debiased accuracy and predicted-position behavior can diverge, and we quantify the scale of position-sensitive score change in a 40-run 3GPP campaign. Third, we show that under matched-budget, permutation-debiased evaluation, support for a data-selection advantage over Random weakens compared to the raw, mismatched-budget setting.

The method-comparison result is specific to the homogeneous 3GPP pool we study; it bounds the reported selector comparison rather than evaluating data-selection methods in general. More broadly, our results support a position-aware reporting protocol for fixed checkpoints, not a model intervention. We release 3GPP-Eval V3.2 (1,220 quality-filtered items), 11,846 domain QA pairs, and code for position-aware reporting.

2 Related Work

Position bias in LLMs. Ko et al. (2020) showed that extractive QA models learn spurious positional cues when the answer-position distribution is skewed. LLM inference also exhibits example-ordering and label sensitivity in few-shot prompting (Zhao et al., 2021); in MCQ settings, reordering options alone can swing accuracy by 13–85% (Pezeshkpour and Hruschka, 2024). Mechanistic

studies further separate content-independent position priors from content-conditioned evidence (Zheng et al., 2024), and localize prediction bias to feed-forward vectors and attention heads (Zhou et al., 2024). In domain and fine-tuning contexts, the evidence is more direct: telecom QA shows measurable position bias that can be mitigated with LoRA (Guda et al., 2025), mechanistic analyses on instruction-tuned models trace positional bias to specific attention heads and transformer layers (Dimino et al., 2025), and fine-tuning itself has been shown to amplify primacy on multi-option classification (Raimondi and Gabbrielli, 2025). These studies establish that option order can affect LLM predictions. They do not, however, directly test whether fine-tuning shifts a checkpoint's position behavior enough to alter comparisons among SFT configurations. We address that gap with a paired MedMCQA contrast and a measurement campaign on 3GPP fine-tuning.

Evaluation reliability. Beyond option order, reported MCQ performance is sensitive to surface and configuration choices: whitespace and label tokens (Sclar et al., 2024), prompt-format variation (Zhuo et al., 2024), and hyperparameter choices (Halfon et al., 2024). Prior work frames these as inference-time prompt effects or training-time hyperparameter effects; less is known about how fine-tuning changes the stability of the evaluation itself. Existing debiasing methods intervene at inference time on model outputs or internal components. PriDe (Zheng et al., 2024) subtracts a position prior from output logits, and UniBias (Zhou et al., 2024) suppresses bias-carrying internal components. We take a complementary reporting approach: rather than modifying logits or internal components, we leave checkpoints unchanged and use permutation-debiased scoring to test whether position effects alter fixed-checkpoint comparisons.

Data selection for instruction tuning. Data selection is a consequential setting for this audit because selector claims are often supported by MCQ benchmark gains. Existing selectors use proxy signals such as instruction informativeness (Li et al., 2024b,a), T-SHIRT-style difficulty (Fu et al., 2025), downstream task alignment (Xia et al., 2024), and base-model uncertainty (Ankner et al., 2025). Large-scale studies present mixed evidence: many selectors do not consistently outperform random sampling in broad instruction-tuning experiments (Ivison et al., 2025), whereas perplexity comparisons at scale find perplexity-based scoring to be a strong predictor of SFT effectiveness (Harada et al., 2025). Those studies mostly use heterogeneous, multi-source instruction corpora. Domain-specific pools derived from 3GPP telecommunications standards (Maatouk et al., 2024; Zou et al., 2025; Kumar et al., 2025) are smaller and more homogeneous; the quality- and diversity-based intuitions that motivate many selectors have been questioned in SFT settings (Shen, 2024). We therefore use these selector families as a method-comparison case rather than ranking selectors per se. The question is whether support for an advantage over the Random baseline changes once training budgets are matched and position-aware reporting is applied.

Table 1: **Evaluation benchmarks and baseline Qwen3-4B accuracies used in the position-aware reporting audit.**

Benchmark	Size	Purpose	Base (%)[†]
L1a: 3GPP-Eval V3.2	1,220	Primary	76.4 / 83.9
L1b: TeleQnA-Std	2,997	External	59.3
L2: TeleQnA-Full	1,992[‡]	General	68.0 / 66.9
L3: MMLU	∼14k	Retention	67.5

[†] Two values indicate raw / permutation-debiased accuracy; L1b and L3 anchor external and retention checks at raw scale. [‡] Stratified subset of the 10,000-item TeleQnA pool, fixed across all runs.

3 Audit Design

We structure the audit around a shared evaluation protocol with three elements: a benchmark whose correct-answer positions are roughly uniform, a position-averaged score that factors out option-order effects, and diagnostics that reveal how each checkpoint distributes predictions across option positions. Each of the four subsequent experiments applies this protocol to a distinct question: whether accuracy and predicted-position behavior can diverge (§4), how large the position-sensitive fraction of accuracy change is (§5), whether position-aware reporting alters support for a method-level conclusion (§6), and whether these patterns generalize across architectures (§5.4).

3.1 Data and Position-Aware Evaluation

Training data. We partition the TSpec-LLM 3GPP corpus (Nikbakht et al., 2024) by specification ID into training and evaluation splits with 0% document overlap (Appendix I). From the training split, we generate 11,846 open-ended

QA pairs with Qwen3.5-Plus[1] and filter them through NLI entailment, terminology matching, and citation validation. Because all selector arms share the same generator and evaluation pipeline, generator–evaluator family effects are controlled at the pipeline level, though selected subsets may still differ in how they interact with such effects. The resulting pool is single-source and narrow in question-type range, defining the homogeneous-pool setting for the selector comparisons below (Ivison et al., 2025).

Evaluation benchmarks. Evaluation uses four benchmarks at two scope levels (Table 1). L1a (3GPP-Eval V3.2, 1,220 MCQ from the held-out split) anchors the primary position-stability analysis; the base Qwen3-4B already shows a 7.5 pp raw-to-debiased gap on this set. L1b (TeleQnA Standards) and L2 (a stratified TeleQnA Full subset) (Maatouk et al., 2026) supply external and generalization checks, while MMLU (Hendrycks et al., 2021) provides a coarse retention signal. As the primary target, L1a passes six quality gates that remove items solvable by lexical overlap, answer-length cues, or skewed answer-position patterns, leaving a near-uniform correct-answer distribution (Appendix I).

Position-averaged scoring and position diagnostics. All MCQ results use option-token log probabilities. For a fixed-order MCQ score to support configuration comparison, the conclusion should not depend on option order; we refer to this requirement as position invariance. Our position-averaged counterpart is *permutation-debiased accuracy*: for each question, $n_{\text{perms}}{=}4$ random option orderings are evaluated, option-token log probabilities are aligned by answer option across permutations and averaged, and the final prediction is taken from the averaged scores. The same approximation is used for both arms in each McNemar test. For the matched comparisons, the $n{=}4$ estimate underestimates full-enumeration accuracy by 2.6–3.9 pp; permutation-seed variation is at least $5\times$ smaller than this gap (Appendix A.1, A.2).

We also report the predicted-position distribution, since two checkpoints can have similar debiased accuracy but different predicted-position distributions. MaxPred% is the largest share in the predicted-position distribution; for a four-option benchmark, a uniform predicted-position distribution has a 25% reference value, and we report deviations from this reference as MaxPred% excess. Directional summaries, such as the Σ statistic introduced for L1a, are used only when a benchmark exhibits a clear dominant drift direction (§5.2).

Statistical testing. All pairwise comparisons use McNemar's exact test (McNemar, 1947) with Bonferroni correction within pre-specified families. The primary family contains 12 IFD-vs-Random tests (3 architectures $\times$ 4 evaluation sets: L1a raw, L1a debiased, L1b raw, L2 raw). This is the correction family for the focal IFD-vs-Random method-comparison claim. We single out this contrast because IFD shows the largest default-reporting advantage over Random in this pool—consistent with the gap reported in the original IFD evaluation (Li et al., 2024b)—making it the comparison most exposed to revision under matched-budget, position-aware controls. Matched comparisons for the remaining selector families are reported separately as supporting context under the same pairwise testing protocol. A broader 36-test family that adds Base-vs-SFT comparisons is reported in Appendix E. Effect sizes are Cohen's h (Cohen, 1988); accuracy estimates include bootstrap 95% confidence intervals (CIs; 10,000 resamples).

3.2 Models and Training

The primary campaign comprises 40 Qwen3-4B (Yang et al., 2025) LoRA runs on the homogeneous 3GPP pool, varying selection method, selection ratio, learning rate, and LoRA rank. The full sweep is used to measure the scale and structure of position-sensitive score change, while method-comparison claims are restricted to runs sharing a single reference training configuration (LoRA rank $r{=}16$, learning rate $2\mathrm{e}{-5}$, 10 epochs) at the same selection ratio, which we call *matched*. Cross-architecture checks use Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024). The primary Qwen inventory and supplementary seed runs are detailed in Appendix B; format-ablation runs are analyzed separately in Appendix H. The method-comparison case covers five selector families that span several recent proxy-scoring approaches for instruction-tuning data selection: IFD (Li et al., 2024b), IFD+FLA (Li et al., 2024a), PPL (Ankner et al., 2025), RDS+ (Ivison et al., 2025), and T-SHIRT (Fu et al., 2025). Selection ratios span 2–100% in the Qwen campaign. Seed replications are reported

[1] https://qwen.ai/blog?id=qwen3.5

in Appendix A.5; released artifacts for the matched IFD-5% vs. Random-5% comparison are listed in Appendix M.

4 Accuracy and Predicted-Position Behavior Can Diverge

We use MedMCQA (Pal et al., 2022) for this diagnostic contrast because each item provides both a standard option layout and an expert-written explanation, allowing the same source to be supervised with or without option-letter exposure (Appendix J.4). The 3GPP corpus does not provide an aligned explanation field, so constructing one would conflate a supervision-format change with a content-generation change.

The contrast fine-tunes Qwen3-4B on 533 MedMCQA items under two supervision formats sharing the same LoRA configuration (r=16, lr=2e−5, 5 epochs). The *Open-QA* variant is trained to generate the expert explanation from the question alone; the *MCQ* variant is trained to first select the correct option letter and then generate the same explanation. Both variants are evaluated on the same 1,460-item MedMCQA-Eval under the default MCQ setting and its permutation-debiased counterpart.

As summarized in Figure 1(a), the two variants show no significant difference in permutation-debiased accuracy (51.8% each, McNemar p=1.00); raw accuracy is also similar (51.7% vs. 51.2%, p=0.70). Under the default MCQ evaluation, however, the MCQ-trained variant shifts MaxPred% by +18.3 pp (28.0 → 46.3%), while the Open-QA variant shifts it by only +3.5 pp—a 5.2× gap in how far probability mass moves toward a preferred position, with no corresponding gap in permutation-debiased accuracy.

The paired contrast establishes an operational separation: the two variants agree in permutation-debiased accuracy but differ in predicted-position distribution. All training inputs, LoRA configuration, and evaluation protocol are shared; the only difference is whether the supervision format includes option letters. The probability-mass shift is therefore associated with option-letter exposure, though the contrast also changes token budget, so we treat it as evidence of separation rather than proof of mechanism. This separation motivates two target-domain questions: how large is the position-sensitive score change on 3GPP (§5), and does it change method comparisons (§6)?

5 Measuring Position-Sensitive Score Change

Having established that accuracy and predicted-position behavior can separate (§4), we now measure the scale of this separation in the target domain: a 40-run Qwen3-4B campaign on 3GPP data, supplemented by cross-architecture checks on Mistral and LLaMA (§5.4). The analysis proceeds from predicted-position concentration (§5.1) to a position-level accuracy model (§5.2), raw-versus-debiased instability (§5.3), and cross-architecture checks.

5.1 Domain SFT Shifts Predicted-Position Preferences on L1a

Domain SFT systematically increases predicted-position concentration on L1a. We measure this concentration by MaxPred% excess, the deviation of the most-predicted position share from the uniform 25% reference. Across all 40 runs, this excess grows roughly 2.6-fold after SFT, with 36 of 40 runs exceeding the base-model level. The dominant drift points toward Opt 1 in 33 of 40 runs. Accuracy change is also position-dependent: Opt 4-correct questions lose 17.6 pp on average, while Opt 1-correct questions gain 7.3 pp. Together with the concentration-tracking checks in Appendix A.3, this pattern motivates the position-level test below: whether stronger Opt 1 drift predicts where accuracy is gained or lost.

5.2 First-Option Drift Predicts Position-Specific Accuracy Change

MaxPred% excess captures how concentrated the selected positions become, but not the direction of the drift. Because L1a drift points toward Opt 1 in 33 of 40 runs (§5.1), we take Opt 1 selection probability as a directional proxy for this benchmark and define Σ as its mean per-question increase after SFT. The question is whether runs with stronger first-option drift gain more on Opt 1-correct questions and lose more on later-position-correct questions. Regressing each position's accuracy change on Σ yields a linear model:

$$\Delta\mathrm{Acc}(k) = \alpha_k \, \Sigma + \beta_k, \qquad (1)$$

where $\Delta\mathrm{Acc}(k)$ is the accuracy change for questions whose correct answer lies at position k ($k \in \{1, 2, 3, 4\}$), and α_k, β_k are position-specific coefficients. Fitted by ordinary least squares over the 40 runs (Figure 2; Table 2), the slopes are positive for

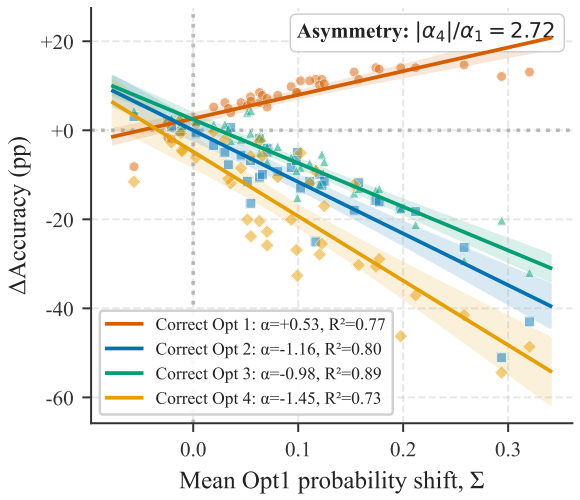

Figure 2: **Position-specific accuracy change versus first-option drift on L1a (40 Qwen3-4B runs).** Each point is one run's ΔAcc for questions whose correct answer is at the indicated position. Σ: mean Opt 1 selection-probability increase after SFT. Lines are OLS fits (Eq. 1); bands show 95% CIs. Stronger Opt 1 drift predicts gains on Opt 1-correct questions and steeper losses at later positions ($|\alpha_4|/\alpha_1$=2.72).

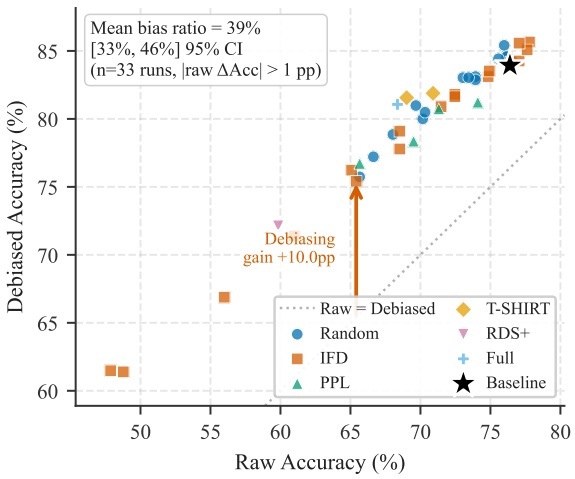

Figure 3: **Raw versus debiased L1a accuracy (40 runs + base): roughly two-fifths of raw change does not persist under option-order averaging.** Points above the diagonal have higher raw than debiased accuracy. The mean gap ratio $|\Delta_{\text{raw}} - \Delta_{\text{debiased}}| \,/\, |\Delta_{\text{raw}}|$ across runs with $|\Delta_{\text{raw}}| \geq 1$ pp is 39% (95% CI [33%, 46%]).

Table 2: **Fitted coefficients for Eq. 1 (40 Qwen3-4B runs).** Positive α_1 and negative $\alpha_{2\text{–}4}$ confirm asymmetric position-specific effects. All bootstrap 95% CIs exclude zero. LOO MAE = 3.7 pp.

	Opt 1	Opt 2	Opt 3	Opt 4
α_k	$+0.53$	-1.16	-0.98	-1.45
β_k	$+0.03$	-0.00	$+0.02$	-0.05
R^2	0.77	0.80	0.89	0.73

Opt 1 and negative for Opt 2–4, with the steepest loss at Opt 4. The ratio $|\alpha_4|/\alpha_1 = 2.72$ summarizes the core asymmetry: Opt 1 drift is associated with a much larger loss on Opt 4 than gain on Opt 1. Out-of-sample and influence diagnostics support the same structure (Appendix G, Appendix A.4).

Because probability mass is redistributed rather than created, the four slopes partially offset one another; their sum is negative, consistent with a net accuracy cost associated with Opt 1 drift. The directional pattern replicates across architectures (§5.4).

5.3 Much of the Raw Accuracy Change Is Position-Sensitive

The position-dependent model predicts that permutation-debiased scoring should reduce fixed-order gains more strongly in higher-concentration runs. The observed pattern is consistent: multi-permutation scoring (n_{perms}=4; §3.1) reduces systematic position concentration, with MaxPred% dropping from 29.8–68.2% to 26.4–28.1% across runs. The raw-to-debiased gain scales with position concentration ($r = +0.75$, $p<0.001$), and among configurations at the 4–5% selection ratio, the apparent raw spread narrows sharply (Figure 3; convergence and validity checks in Appendix A.1, A.3).

We summarize the scale with the gap ratio $|\Delta_{\text{raw}} - \Delta_{\text{debiased}}| \,/\, |\Delta_{\text{raw}}|$ (each Δ relative to the base; operational, not causal). Across runs with $|\Delta_{\text{raw}}| \geq 1$ pp, the mean ratio is 39% (bootstrap 95% CI [33%, 46%]): roughly two-fifths of raw accuracy change does not persist under option-order averaging.

5.4 The Directional Pattern Repeats Across Architectures

The same directional structure appears outside Qwen: fitting Eq. 1 independently on Mistral-7B-Instruct-v0.3 (N=9) and LLaMA-3.1-8B-Instruct (N=7) yields positive α_1 and negative $\alpha_{2\text{–}4}$ in both cases (Table F1). Magnitudes remain architecture-specific, so these fits confirm the sign pattern rather than a shared numerical model; the smaller samples preclude re-estimation of the gap ratio (Appendix F). The next question is whether this option-order instability changes a method comparison.

6 Position-Aware Reporting Alters a Method Comparison

Having measured position-sensitive score change on 3GPP (§5), we next ask whether it changes the support for a method-level comparison. Data selection is a useful test case because small MCQ accuracy gaps are often used to choose among fine-tuning configurations, and the analysis tests whether the reporting controls alter such a comparison rather than ranking selectors.

6.1 Matched and Position-Aware Comparison

The comparison adds two controls motivated by the preceding analysis: multi-permutation scoring for position-sensitive score changes and matched hyperparameters for training-budget differences. With both controls in place, the observed matched gaps are small and do not survive the pre-specified correction. At the 5% ratio on L1a, the matched debiased IFD–Random gap is 1.64 pp (p_{Bonf}=0.182, $|h|$=0.042; k=12). Other matched 5% selector runs show similarly small debiased gaps over Random (Figure 1b; Appendix E).

This small-gap pattern is not confined to the 5% matched point. At nearby ratios and on the external L1b check, matched selector gaps remain small and non-significant (Appendix E, Appendix C). This non-detection is consistent with the single-source pool setting; §6.2 shows that selector gaps re-emerge under mixed-domain conditions.

Seed-level replication establishes the detection threshold: the IFD 5% point estimate lies within the Random 5% seed variation, and the current design cannot rule out advantages below the detection threshold (2.4 pp on L1a; Appendix A.5). Because several selector families have only single-run matched coverage, the matched evidence for those families is correspondingly limited.

How the controls change the comparison. Figure 4 shows why both controls matter: budget matching accounts for the larger reduction (from +4.43 to +2.30 pp on L1a), and permutation-debiased scoring narrows the gap further to +1.64 pp (n.s.). Separately, changing only LoRA rank (r=16→64 at 4% data) shifts MaxPred% excess by +12.6 pp, exceeding the observed selector differences.

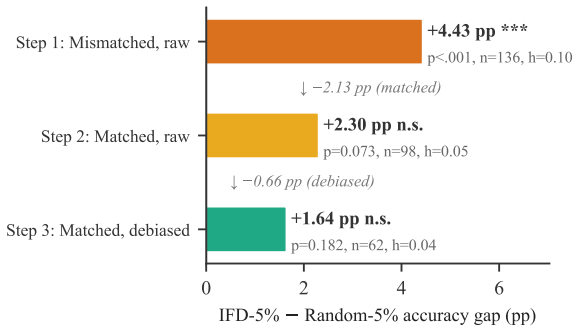

Figure 4: **IFD-5% vs. Random-5% L1a gap under progressively stricter controls.** The gap falls from +4.43 pp (raw, mismatched) to +2.30 pp (matched) to +1.64 pp (matched + debiased; p_{Bonf}=0.182, n.s.; k=12). Bar annotations: Bonferroni-corrected p, McNemar discordant-pair count n, Cohen's $|h|$.

6.2 Selector Gaps Re-Emerge in Heterogeneous Pools

The small matched selector gaps in the homogeneous 3GPP-pool setting may reflect pool composition rather than a broad property of data-selection methods. The 11,846 training samples come from a single specification family (3GPP), share one generation pipeline, and span a narrow question-type range. We therefore test the boundary with a 50% 3GPP + 50% Alpaca mixed pool under identical LoRA hyperparameters (Appendix J).

Selector gaps then re-emerge in this heterogeneous follow-up: across four replicate seeds in the 50/50 mixed pool, the IFD–Random debiased L1a gap ranges from +2.0 to +9.9 pp, and the RDS+–Random gap from +1.8 to +6.6 pp. This widening is associated mainly with lower Random performance in the mixed pool (Appendix J.2). Because the correction family and discordance patterns differ from the homogeneous-pool test, we treat this as boundary evidence rather than a direct comparison with the 2.4 pp detection threshold; on L1b and L2, gaps remain below 1.5 pp.

Appendix J.2 confirms that the Eq. 1 directional structure remains descriptive in these mixed-pool runs and reports the bridge decomposition and ratio sweep. Together, these results suggest that pool composition is one boundary condition for the homogeneous-pool finding: in both settings, support for a selector comparison depends on whether raw accuracy is accompanied by position-aware reporting.

7 Scope and Reporting Recommendation

Two practical questions remain: whether the preferred position generalizes across benchmarks, and how practitioners should report MCQ results when comparing fine-tuned checkpoints.

Position effects are benchmark-specific. The preferred position is benchmark-specific rather than model-wide: after 3GPP SFT, Opt 1 selection on L1a rises from 31.7% to 50.7%, whereas on MMLU it falls from 20.2% to 17.4% (Table J6); MedMCQA favors Opt 2 under medical SFT (Appendix J.4). A position check should therefore accompany each evaluation set, not just each model.

The position-aware reporting protocol. For each benchmark used in a configuration comparison, we recommend reporting raw accuracy, permutation-debiased accuracy, and the predicted-position distribution. Raw accuracy reports the conventional score; permutation-debiased accuracy tests stability to option order; and the predicted-position distribution shows whether reported scores are accompanied by different position behavior. When training configurations are compared, the same report should include significance tests under matched training budgets and state the pre-specified correction family.

When most drift concentrates on one position, the directional model (Eq. 1) separates a run's raw accuracy change into position-predicted and residual components; otherwise, the full predicted-position distribution serves as the more general summary. Cross-domain controls in Appendix J illustrate these outputs on non-3GPP adapters. This protocol is not a model intervention: it operates on option-token log probabilities from fixed checkpoints, requiring no retraining or free-form answer generation.

8 Discussion and Conclusion

Raw MCQ accuracy remains useful, but it is incomplete evidence for configuration-level claims when fine-tuned checkpoints differ in position behavior. The central implication is a measurement one: fixed-order scores should not be read alone when the conclusion may depend on where the correct option appears.

The evidence follows three steps: the MedMCQA contrast establishes separability, the 3GPP campaign measures target-domain scale, and the method-comparison case shows that matched-budget, position-aware reporting can change support for the tested IFD data-selection method over Random on a homogeneous pool. Because the preferred position varies with the benchmark rather than the model (§7), neither an architecture-specific nor a model-wide primacy explanation fits the evidence.

Candidate mechanisms. The evidence is consistent with format exposure and probability-mass redistribution, but does not isolate a single causal pathway. This aligns with observations that fine-tuning can behave partly as format adaptation (Ovadia et al., 2024; Gudibande et al., 2024; Zhou et al., 2023); supporting ablations and probes remain observational (Appendix H, D, K). Importantly, the reporting recommendation does not depend on resolving the mechanism; it follows from the observed instability of fixed-order comparisons. Testing the mechanism directly would require targeted component interventions (Dimino et al., 2025).

Scope of the recommendation. For coarse model selection or cross-version regression tests, raw MCQ accuracy remains a useful quick signal. The recommendation is narrower: position behavior can affect the comparison once compared configurations differ in format exposure or training budget, such as a domain-SFT checkpoint against its base or two selected subsets trained under different schedules. The §7 reporting protocol adds this check without retraining or modifying the checkpoint, although it does require additional log-probability evaluations.

Takeaways and open questions. Our results suggest that MCQ reports used for domain-SFT configuration comparison should include position checks. In the primary campaign, permutation-debiased scoring makes visible a large raw-to-debiased gap that fixed-order accuracy alone would hide. We release 3GPP-Eval V3.2, 11,846 QA pairs, and the reporting code of §7 in an anonymized repository.[2] Two questions follow: whether training- or prompt-side changes can restore approximate position invariance without relying on permutation-debiased scoring, and whether similar format-related position effects arise in other evaluation formats, such as open-ended generation, pairwise preference, or span extraction.

[2] https://anonymous.4open.science/r/3GPP_MCQ-7AB4

Limitations

Scope. The primary campaign covers one domain (3GPP telecommunications) with same-family data generation (Qwen3.5-Plus → Qwen3-4B), so the effect magnitude could partly reflect bias from this shared model family. Cross-architecture replications on Mistral and LLaMA provide a partial check, and the MedMCQA paired contrast (Appendix J.4) provides an independent diagnostic setting. A similarly large-scale replication across generator families and domains remains future work. All evaluation is MCQ-based, leaving effects on open-ended generation untested. Training uses LoRA throughout; full-parameter fine-tuning could interact differently with positional representations. The heterogeneous-pool follow-up covers one domain mixture (3GPP + Alpaca) at three proportions; whether the compositional effect generalizes to other domain combinations is not yet known. These choices limit claims about effect magnitude across domains, model families, and evaluation formats, while the within-setting separation between position-averaged accuracy and predicted-position distribution is directly supported by the matched evaluation reported here.

Model coverage. Cross-architecture regression on Mistral-7B-Instruct (9 runs) and LLaMA-3.1-8B-Instruct (7 runs) reproduces the slope asymmetry, though with smaller samples and a weaker fit for Mistral Opt 4. The MedMCQA paired contrast uses one training seed per variant. It supports a separation between permutation-debiased accuracy and predicted-position distribution in that tested contrast, but it does not estimate seed-level variability. The primary model (Qwen3-4B) is a base checkpoint with strong MCQ performance. This places it in a shared fine-tuning regime with the instruction-tuned Mistral and LLaMA variants used for cross-architecture checks, since the corresponding base checkpoints score near chance ($\sim$26%). Instruction tuning may, however, introduce position priors absent from a base model, so cross-architecture magnitude differences could partly reflect this asymmetry rather than purely architectural variation. Separately, three of five selector families (PPL, RDS+, IFD+FLA) have only single-run coverage at the matched point. The McNemar power floor (§6.1) leaves advantages below 2.4 pp undetectable, so the homogeneous-pool result is consistent with both no selector advantage and small advantages that fall below the detection threshold.

Measurement. The head-level probing (Appendix K) covers one architecture and six runs, yielding a correlational rather than causal link between attention patterns and Σ. The Σ statistic is an L1a-specific directional summary, appropriate because most Qwen3-4B runs drift toward Opt 1 on that benchmark. Benchmarks with different preferred positions require their own summaries or the full predicted-position distribution. The n_{perms}=4 debiasing budget underestimates the full-enumeration debiased score (Appendix A), consistently in the conservative direction for measured position-sensitive share; full enumeration would sharpen the magnitude estimates at higher inference cost. The MedMCQA paired contrast changes both option-letter supervision and token budget, so isolating those factors would require a dedicated three-way format ablation. External item-level annotations would further separate content validity from position-sensitive scoring behavior. All MCQ scores are computed from per-option log-probabilities; alternative protocols such as free-generation parsing or full-option-text scoring may exhibit different position-preference profiles.

Ethics Statement

All data used in this study—3GPP technical specifications, open-weight language models, and the TeleQnA benchmark—are publicly available. No human subjects or private data were involved, so institutional review board approval was not required.

Training QA pairs were generated by Qwen3.5-Plus from public specification text and then filtered with three automated checks: NLI entailment, terminology matching, and citation validation (§3.1). These filters reduce, but cannot eliminate, residual factual errors; the released training data are therefore explicitly labeled as machine-generated. The 3GPP-Eval V3.2 benchmark is derived entirely from publicly distributed standards documents and contains no proprietary or access-restricted material.

We release the 3GPP-Eval V3.2 benchmark, the 11,846 training QA pairs, and the reporting code under CC-BY-4.0 to support reproducibility. The benchmark is intended for studying position effects in domain SFT, not as a certification of telecommunications competence.

The full study includes 121 LoRA fine-tuning experiments spanning six campaigns. Appendix B documents the primary 3GPP inventory; other sup-

plementary run groups are introduced in the appendix sections where they are analyzed. All runs that completed are included in the analysis pools declared for each comparison; no run was excluded post hoc on the basis of its outcome. Including the associated multi-permutation evaluation on four benchmark suites, this amounted to approximately 250 GPU-hours on a single NVIDIA A100-40 GB (§3.2).

One concrete misuse risk is selective reporting of whichever metric (raw or debiased) favors a particular method; we therefore recommend reporting both scores together with the predicted-position distribution, allowing readers to judge the position-sensitive part of the comparison directly.

References

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max M. Marion, Matthew Leavitt, and Mansheej Paul. 2025. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In *Proceedings of the Thirteenth International Conference on Learning Representations*, pages 100615–100634.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Routledge.

Fabrizio Dimino, Krati Saxena, Bhaskarjit Sarmah, and Stefano Pasquali. 2025. Tracing positional bias in financial decision-making: Mechanistic insights from Qwen2.5. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 96–104.

Yanjun Fu, Faisal Hamman, and Sanghamitra Dutta. 2025. T-SHIRT: Token-selective hierarchical data selection for instruction tuning. In *Advances in Neural Information Processing Systems 38*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Blessed Guda, Lawrence Francis, Gabrial Zencha Ashungafac, Carlee Joe-Wong, and Moise Busogi. 2025. Quantifying and mitigating selection bias in LLMs: A transferable LoRA fine-tuning and efficient majority voting approach. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, pages 2019–2038, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2024. The false promise of imitating proprietary language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Alon Halfon, Shai Gretz, Ofir Arviv, Artem Spector, Orith Toledo-Ronen, Yoav Katz, Liat Ein-Dor, Michal Shmueli-Scheuer, and Noam Slonim. 2024. Stay tuned: An empirical study of the impact of hyperparameters on LLM tuning in real-world applications. *arXiv preprint arXiv:2407.18990*.

Yuto Harada, Yusuke Yamauchi, Yusuke Oda, Yohei Oseki, Yusuke Miyao, and Yu Takagi. 2025. Massive supervised fine-tuning experiments reveal how data, layer, and training factors shape LLM alignment quality. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22360–22381, Suzhou, China. Association for Computational Linguistics.

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

Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. 2025. Large-scale data selection for instruction tuning. *arXiv preprint arXiv:2503.01807*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1121.

Anshul Kumar, Gagan Raj Gupta, Manish Rai, Apu Chakraborty, Ashutosh Modi, Abdelaali Chaoub, Soumajit Pramanik, Moyank Giri, Yashwanth Holla, Sunny Kumar, and M. V. Kiran Sooraj. 2025. MM-Telco: Benchmarks and multimodal large language models for telecom applications. *arXiv preprint arXiv:2511.13131*.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024a. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14255–14273.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024b. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 7602–7635.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025. Preserving diversity in supervised fine-tuning of large language models. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Ali Maatouk, Kenny Chirino Ampudia, Rex Ying, and Leandros Tassiulas. 2024. Tele-LLMs: A series of specialized large language models for telecommunications. *arXiv preprint arXiv:2409.05314*.

Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Merouane Debbah, and Zhi-Quan Luo. 2026. TeleQnA: A benchmark dataset to assess large language models telecommunications knowledge. *IEEE Network*, 40(2):253–260.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Rasoul Nikbakht, Mohamed Benzaghta, and Giovanni Geraci. 2024. TSpec-LLM: An open-source dataset for LLM understanding of 3GPP specifications. In *2024 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico.

Bianca Raimondi and Maurizio Gabbrielli. 2025. Exploiting primacy effect to improve large language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, pages 989–997.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. 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*, pages 54104–54132.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems 36*.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024. UniBias: Unveiling and mitigating LLM bias through internal attention and FFN manipulation. In *Advances in Neural Information Processing Systems 37*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976.

Hang Zou, Qiyang Zhao, Yu Tian, Lina Bariah, Faouzi Bader, Thierry Lestable, and Merouane Debbah. 2025. TelecomGPT: A framework to build telecom-specific large language models. *IEEE Transactions on Machine Learning in Communications and Networking*, 3:948–975.

Table A1: **Debiased L1a accuracy (%) at three permutation budgets.** All evaluations use permutation seed 42 on 1,220 L1a v3.2 questions. The n=24 column is exhaustive (4!); n=4 and n=8 sample from the same permutation space. † High-concentration rows use a consistent LoRA-adapter loading protocol; the top three rows follow the main-analysis merged-model evaluation protocol.

Configuration	n=4	n=8	n=24	$\Delta_{4\to24}$	$\Delta_{8\to24}$
Baseline	83.9	84.7	86.6	+2.6	+1.9
IFD 5% (2e−5)	81.6	81.5	85.6	+3.9	+4.1
Random 5% (2e−5, 10ep)	80.0	80.7	83.6	+3.6	+3.0
IFD 10% (2e−5)†	61.2	63.8	70.2	+8.9	+6.4
IFD 15% (2e−5)†	60.0	62.0	69.0	+9.0	+7.0
IFD 20% (2e−5)†	60.4	62.3	71.3	+10.9	+9.0

A Permutation-Debiasing Validation

A.1 Convergence: $n_{\text{perms}} = 4$ vs. 8 vs. 24

The main analysis sets n_{perms}=4 because full enumeration (4!=24) would multiply inference cost sixfold across all runs and 1,220 evaluation questions. Table A1 quantifies the resulting approximation error for six configurations spanning the full observed position-concentration range.

For the three configurations used in the main paired comparisons (raw MaxPred 25–45%), the n=4→n=24 gap is 2.6–3.9 pp. For the high-concentration dose-response tail (IFD 10–20%, raw MaxPred >60%), the gap widens to 8.9–10.9 pp. The growth is expected: debiased accuracy is an argmax over aggregated log-probabilities, so convergence depends on the margin between leading options; when raw prediction mass concentrates more strongly on the preferred position, each permutation reorders a wider portion of the distribution, and more permutations are needed for the aggregate to settle.

Two features limit the practical impact of this underestimation. First, the direction is uniform across all six configurations: n=24 always yields higher debiased accuracy than n=4, so the approximation underestimates the position-averaged score. The 39% raw-to-debiased gap ratio reported in §5.3 is conservative on this axis as well: the n=4 approximation deflates the debiased score and thus deflates the measured share of raw change that is not stable under permutation-debiased scoring. Second, the IFD–Random comparison is stable across the two permutation budgets: the debiased gap is 1.64 pp at n=4 and 1.96 pp at n=24 (small in both cases, with IFD slightly above Random), with both conditions sharing the same permutation budget and seed.

A.2 Seed Stability

Across permutation seeds 42, 123, and 456, debiased accuracy varies by 0.17–0.47 pp, at least 5× smaller than the n=4→24 convergence gap (2.6–10.9 pp; §A.1) and well below the 7–16 pp debiasing effects analyzed in the main text. The choice of seed 42 does not materially affect any reported comparison.

A.3 Indirect Validity Checks

Convergence and seed stability establish that the debiasing estimate is numerically reproducible. A separate question is whether the position-averaged score is less entangled with option position than the raw fixed-order score. Three structural patterns in the data bear on this.

Position-concentration tracking. If permutation-debiased scoring introduced its own position-unrelated shift, the gain should not track independently measured position concentration. In practice, the correlation between debiasing gain and MaxPred% excess is $r = +0.75$ (p<0.001) across the 40 primary runs and the base model. Runs with MaxPred% near 25% (i.e., near-uniform prediction distributions) show gains below 1 pp, while runs with MaxPred% >50% recover 10–14 pp. This monotone relationship is expected when the raw fixed-order score contains a position-related component.

Option-count dilution. L2 (TeleQnA-Full) contains 64.4% five-option questions and uses a different prompt template. Maximum debiasing gain on L2 is 3.83 pp (IFD 5%), versus 13.6 pp on four-option L1a. Under a single-position preference, moving from four to five options dilutes the position-dependent component geometrically. The observed gain ratio is roughly consistent with this expectation. A position-unrelated score shift would not produce this scaling.

Cross-domain controls. The Alpaca and medical-domain controls (§J) produce MaxPred% within ±4 pp of the base model and show debiasing gains below 1.5 pp. The same scoring procedure applied to 3GPP runs recovers 7–14 pp. If permutation-debiased scoring systematically inflated scores regardless of predicted-position behavior, the non-3GPP controls should show comparable gains; they do not.

External item-level annotations of content validity would provide stronger confirmation. Nevertheless, these three checks narrow the space of alternative explanations: they weigh against the hypothesis that debiasing gain is an additive constant or noise-driven effect, and they are consistent with a scoring procedure that averages over option-position sensitivity.

A.4 Robustness of the Position-Dependent Model

Cook's distance analysis identifies IFD 15% and IFD 20% as high-leverage points for most positions (Cook's $D > 4/n = 0.08$); both lie in the extreme tail of the Σ distribution ($\Sigma > 0.29$). Excluding all influential points (2–4 per position), the re-fitted slopes retain the same sign and asymmetric structure: $\Delta\alpha_k$ is $+14\%$, $+23\%$, -1%, and -5% for Opt 1–4 respectively. The direction and ordering of the slopes are stable; only Opt 1 and Opt 2 magnitudes shift appreciably, driven by removal of the high-concentration tail where accuracy changes saturate. A Breusch-Pagan test rejects homoscedasticity only for Opt 2 (BP=8.47, p<0.01); the remaining positions show no significant residual-variance dependence on Σ (p>0.05). The Opt 2 heteroscedasticity is consistent with a floor effect at high Σ, where Opt 2 accuracy has less room to decline. The qualitative conclusions in the main text hold across these checks.

A.5 Data-Selection Seed Variance

Random 5% was replicated with six data-sampling seeds (42, 123, 456, 789, 1024, 2025) under the matched lr=2e−5 / 10-epoch configuration, yielding debiased L1a from 78.03% to 82.95% (σ = 1.89 pp). A separate seven-seed replication at lr=5e−5 / 5 epochs yields σ = 3.0 pp (range 74.34–83.11%), consistent with the expectation that higher learning rates amplify seed-to-seed variability. Four supplementary lr=5e−5 seeds († rows in Table B1) are excluded from the 40-run primary analysis to keep the comparison matrix balanced.

IFD 5% and T-SHIRT 5% were each replicated with four training seeds (42, 123, 456, 789) under the same matched configuration. IFD debiased L1a ranges from 81.64% to 82.05% (σ = 0.18 pp); T-SHIRT from 80.49% to 81.56% (σ = 0.53 pp). IFD variability is an order of magnitude below Random's σ = 1.89 pp, and T-SHIRT variability is also substantially smaller ($\sim$3.6×), indicating that selector-side training noise is well below data-sampling noise in these replications. MaxPred% is similarly stable within each selector: IFD σ_{MaxP} = 0.66 pp (range 42.1–43.7%), T-SHIRT σ_{MaxP} = 1.23 pp (range 38.2–41.1%).

Heterogeneous pool. A parallel four-seed replication was conducted on the domain-stratified heterogeneous pool (50% 3GPP / 50% Alpaca, 533 samples; Appendix J.2). Random was replicated with data-sampling seeds 42, 123, 456, 789; IFD and RDS+ were replicated with training seeds 123, 456, 789 in addition to the original seed 42, holding data selection fixed. Random debiased L1a ranges from 72.3% to 80.8% (σ=3.56 pp), roughly twice the homogeneous-pool variance and consistent with the added instability from domain mixing. IFD ranges from 82.2% to 83.0% (σ=0.35 pp); RDS+ from 78.9% to 82.6% (σ=1.64 pp). The IFD–Random gap is positive at all four seeds (+2.0 to +9.9 pp) and each pairwise McNemar test survives Bonferroni correction within the four-seed family (k=4); the same holds for RDS+–Random (+1.8 to +6.6 pp). Full per-seed scores appear in Table J2; bridge decomposition results are reported in Tables J3–J4.

B Experiment Inventory

Table B1 lists the 40 Qwen3-4B SFT configurations used for the main 3GPP campaign plus 4 supplementary seed replications (†). An additional 11 seed-replication runs used for variance estimation are documented in Appendix A.5; together with the 44 runs below, these form a broader 55-run homogeneous-pool Qwen3-4B set used for supplementary robustness analyses unless a comparison explicitly declares the 40-run primary set. The seven format-ablation runs are analyzed separately in Appendix H. All runs use LoRA on the same 3GPP training pool (11,846 QA pairs); manipulated variables are selector method, selection ratio, learning rate, LoRA rank, epoch budget, and random seed. Rows are sorted by debiased L1a in descending order; shaded rows mark the matched-control comparison points discussed in §6.1. To locate the primary IFD-vs-Random comparison, see rows 19 and 27.

C IFD Dose-Response

Table B1 records L1a accuracy for IFD ratios spanning 2% to 100%; this section reports the L1b counterpart. On L1b, low-ratio IFD runs are stable (2%: 60.03%, 3%: 60.69%, 4%: 61.26%, 5%:

Table B1: **Experiment inventory: 40 primary runs + 4 supplementary seeds (†) + base model.** Sorted by debiased L1a (descending). † rows are excluded from the 40-run quantitative analysis. Shaded rows mark matched-control comparison points (§6.1). The ID column preserves original experiment identifiers; format-ablation runs are reported separately (Appendix H). Db = debiased; MaxP = MaxPred%. Variant labels: SF = Superfiltering IFD computation; d-bal. = difficulty-balanced FLA; PPL-double = type × length-stratified; PPL-low = lowest-PPL; IFD-Strat = question-type-stratified IFD.

ID	Experiment	%	lr	Rank	Ep	Steps	L1a (%)		MaxP	L1b	L2	MMLU
							Raw	Db	%	%	%	%
	Base model	—	—	r16	—	—	76.4	83.9	31.7	59.3	68.0	67.5
1	IFD 2%	2	2e−5	r16	10	140	77.8	85.7	31.3	60.0	67.5	67.2
2	IFD<1 (SF)	<1	2e−5	r16	10	140	77.1	85.6	31.6	60.0	67.5	67.1
3	Random 2% (s123)	2	2e−5	r16	10	140	76.0	85.4	32.3	59.7	67.6	67.1
4	IFD<1	<1	2e−5	r16	10	60	77.6	85.1	29.8	59.7	67.7	67.2
5	Random 2%	2	2e−5	r16	10	140	76.2	84.6	31.8	59.6	67.6	67.0
6	Random 2% (s456)	2	2e−5	r16	10	140	76.2	84.6	30.7	59.7	67.3	67.3
7	Random 5% (ep5)	5	2e−5	r16	5	170	75.6	84.4	33.4	60.2	67.9	67.2
8	IFD 3%	3	2e−5	r16	10	200	77.1	84.3	35.1	60.7	67.2	67.0
9	IFD+FLA 3%	3	2e−5	r16	10	200	76.5	84.2	36.0	61.3	68.1	67.2
10	IFD 4%	4	2e−5	r16	10	270	74.9	83.5	38.5	61.3	68.5	67.1
11	IFD+FLA 4% (d-bal.)	4	2e−5	r16	10	270	74.8	83.1	39.4	61.7	68.0	67.0
12	Random 4%	4	2e−5	r16	10	270	73.9	83.1	39.8	61.2	67.9	67.3
13†	Random 5% (s3407)	5	5e−5	r16	5	170	71.5	83.1	31.6	59.6	67.1	—
14	Random 5% (s456, 5e−5)	5	5e−5	r16	5	170	73.4	83.0	31.8	58.6	66.7	—
15	Random 4% (s456)	4	2e−5	r16	10	270	73.0	83.0	37.2	60.2	67.2	67.4
16	Random 4% (s123)	4	2e−5	r16	10	270	73.9	82.9	38.0	60.7	68.1	67.2
17	T-SHIRT 4%	4	2e−5	r16	10	270	70.9	81.9	39.4	53.1	62.1	66.3
18	IFD+FLA 4%	4	2e−5	r16	10	270	72.5	81.8	40.7	61.2	68.2	66.9
19	IFD 5%	5	2e−5	r16	10	340	72.5	81.6	43.7	60.7	68.6	67.0
20	T-SHIRT 5%	5	2e−5	r16	10	340	69.0	81.6	41.1	59.8	67.5	66.6
21	PPL-double 5%	5	2e−5	r16	10	340	74.1	81.2	35.6	59.9	67.8	67.0
22	Random 5% (s123, 5e−5)	5	5e−5	r16	5	170	69.7	81.0	40.5	59.6	67.5	—
23	IFD-Strat 5%	5	2e−5	r16	10	340	71.5	80.9	44.4	61.1	68.5	66.8
24	PPL-low 5%	5	2e−5	r16	10	340	71.3	80.7	39.0	59.8	67.6	66.6
25	Full pool (100%)	100	5e−5	r16	5	3335	68.4	81.1	49.4	60.7	67.6	65.2
26	Random 5% (r64)	5	5e−5	r64	2	68	70.3	80.5	43.5	60.9	67.8	—
27	Random 5% (ep10)	5	2e−5	r16	10	340	70.2	80.0	45.2	60.5	68.8	—
28	IFD+FLA 4% (r64)	4	2e−5	r64	7	189	68.5	79.1	45.7	61.3	68.0	67.1
29	Random 5% (s42, 5e−5)	5	5e−5	r16	5	170	68.0	78.9	44.3	60.1	67.2	—
30†	Random 5% (s1024)	5	5e−5	r16	5	170	67.3	78.7	45.1	60.6	67.8	—
31	PPL-low 5% (5e−5)	5	5e−5	r16	5	170	69.5	78.4	40.0	57.7	65.0	—
32	IFD 5% (ep15)	5	2e−5	r16	11	374	68.5	77.8	48.9	60.2	67.9	66.5
33	Random 4% (r64)	4	2e−5	r64	7	189	66.6	77.2	46.1	60.4	68.3	66.9
34	PPL-double 5% (5e−5)	5	5e−5	r16	5	170	65.7	76.7	44.0	58.2	65.8	—
35†	Random 5% (s789)	5	5e−5	r16	5	170	64.1	76.6	46.6	60.6	68.3	—
36	IFD 4% (r64)	4	2e−5	r64	7	189	65.1	76.2	51.1	61.7	68.7	67.3
37	Random 5% (r64, 2e−5)	5	2e−5	r64	7	238	65.7	75.7	49.0	61.7	68.9	66.6
38	IFD 5% (5e−5)	5	5e−5	r16	5	170	65.4	75.4	50.7	61.3	69.1	67.0
39†	Random 5% (s2025)	5	5e−5	r16	5	170	63.0	74.3	52.6	62.1	69.3	—
40	RDS+ 5%	5	5e−5	r16	5	170	59.8	72.1	54.8	55.6	62.1	—
41	IFD 5% (5e−5, ep7)	5	5e−5	r16	7	238	61.0	71.4	55.0	61.6	69.1	67.0
42	IFD 10%	10	2e−5	r16	10	670	56.0	66.9	61.2	60.2	67.1	65.9
43	IFD 15%	15	2e−5	r16	9	900	48.8	61.4	68.2	52.8	58.1	65.7
44	IFD 20%	20	2e−5	r16	6	804	47.9	61.5	65.9	48.2	52.1	65.4

60.73%), and performance holds near baseline at 10% (60.19%). However, it drops sharply at 15% and 20% to 52.82% and 48.21% respectively, before recovering to 60.73% at 100%. Both drops are individually significant in post-hoc McNemar tests versus IFD-5%, though they fall outside the pre-specified 12- or 36-test families (Appendix E). The resulting pattern is non-monotonic, consistent with mid-range IFD ratios reducing pool diversity without yielding a reliably stronger supervision signal (Li et al., 2025).

D Answer-Flip Direction Analysis

Across all 40 runs, we track every question that the base model answers correctly but the SFT model answers incorrectly (a correct→wrong flip). Figure D1 shows the flow from the correct-answer position to the SFT-predicted (wrong) position. Of the 5,298 such flips, 76.5% land on Opt 1, the position that gains prediction mass in the L1a analysis (§5.1). The source side is also asymmetric: 44.3% of flips originate from Opt 4-correct questions, the position with the steepest accuracy loss in the position-dependent model ($\alpha_4 = -1.45$; Table 2). The concentration at the Opt 4→Opt 1 cell (30.2% of all flips) provides item-level evidence

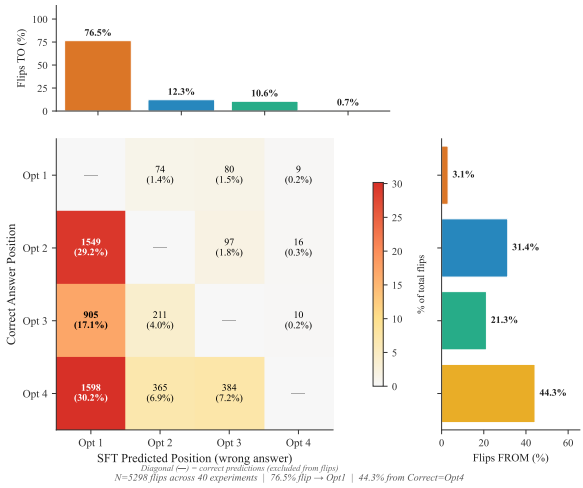

Figure D1: **Correct→wrong flip flow across the 40-run campaign** (N=5,298 flips). Rows: correct-answer position; columns: SFT prediction. Marginal bars summarize target (top) and source (right) distributions. 76.5% of all flips land on Opt 1, and 44.3% originate from Opt 4-correct questions.

that Opt 1 drift is associated with disproportionate errors on later-position questions.

E Cross-Architecture McNemar Results

Two pre-specified test families govern multiple-comparison correction throughout the paper. The **selector family** contains 12 tests: IFD-5% versus Random-5% on each of 3 architectures × 4 evaluation measures (L1a raw, L1a debiased, L1b, L2), with Bonferroni threshold $\alpha' = 0.05/12 \approx 0.004$. The **full cross-architecture family** adds Base-vs-IFD and Base-vs-Random comparisons on the same 3 × 4 grid, yielding 36 tests total ($\alpha' = 0.05/36 \approx 0.001$). All main-text selector conclusions (§6, §5.4) use the 12-test family; the broader 36-test family appears only in this appendix. The 12 IFD-vs-Random results are the primary selector evidence; the remaining 24 Base-vs-SFT comparisons quantify the SFT learning effect itself and are reported below for completeness.

All three architectures show lower debiased L1a accuracy after domain SFT: Qwen 83.93%→80.00% (−3.93 pp), Mistral 69.92%→63.11% (−6.80 pp), LLaMA 75.33%→67.87% (−7.46 pp). The uniform direction is consistent with the homogeneous-pool pattern discussed in §5.4.

Table E1 reports the complete 36-test matrix. Figure E1 plots the IFD-vs-Random effect sizes across the 12 selector-family comparisons. All four

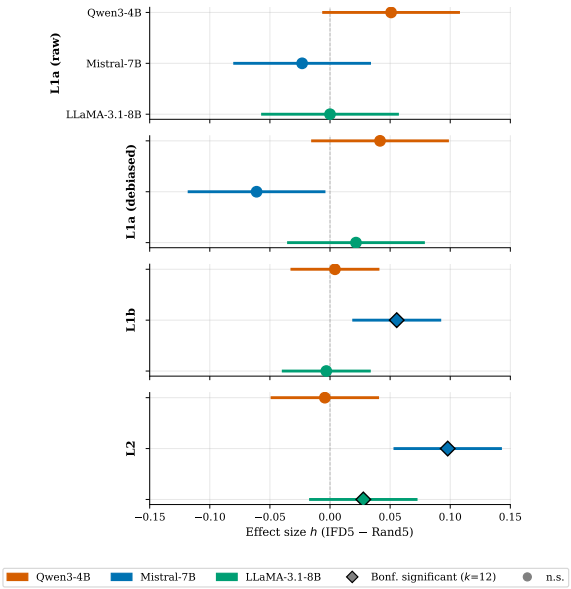

Figure E1: **Forest plot of IFD-vs-Random effect sizes** (Cohen's h) across 3 architectures × 4 evaluation sets. Evaluation sets: L1a (3GPP Layer 1 sub-task A, four-option MCQ) reported in both raw and debiased form; L1b (3GPP Layer 1 sub-task B, four-option MCQ); L2 (3GPP Layer 2, five-option MCQ). Error bars show 95% CIs. Diamond markers denote comparisons that remain significant after Bonferroni correction (k=12). Two LLaMA L1a comparisons reach significance in the negative direction (Random outperforms IFD); all Mistral and Qwen comparisons are non-significant.

Mistral comparisons are non-significant ($|h| \leq 0.03$, $p_{\text{Bonf}} \geq 0.94$). On LLaMA, Random outperforms IFD on raw L1a (−2.21 pp, p_{Bonf}=0.038) and debiased L1a (−2.95 pp, p_{Bonf}=0.004); the remaining ten comparisons are non-significant.

F Cross-Architecture Per-Position Analysis

Section 5.4 reports the cross-architecture regression in summary form. Table F1 summarizes the per-architecture slopes; Table F2 provides the full per-position parameters for all three model families.

On Qwen3-4B, the 55-run appendix fit includes all homogeneous-pool SFT runs. The 40-run primary subset in Table 2 yields tighter fits (R^2_{min}=0.73 vs. 0.61 here): the 15 supplementary and seed-replication runs excluded from that set (Appendix B) add sampling noise without expanding the Σ range.

The shared directional pattern documented in §5.4 masks a within-architecture difference in how accuracy loss is distributed: Opt 4 shows the steep-

Table E1: **Full 36-test McNemar matrix across comparison types, architectures, and evaluation sets.** Δ pp $= \mathrm{Acc}_B - \mathrm{Acc}_A$; discordant pairs: ($n_{A>B}$ / $n_{B>A}$). IFD-vs-Random rows use the 12-test selector-family correction (α'_{12}=0.004); all other rows use the 36-test correction (α'_{36}=0.001).

Comparison	Arch	Eval	N	Disc. ($n_{A>B}$ / $n_{B>A}$)	Δ pp	p_{raw}	p_{Bonf}	h
Panel A: Random-5% (A) vs IFD-5% (B) — selector family, k=12								
	Qwen	L1a raw	1220	35 / 63	+2.30	.006	.073	+.051
	Qwen	L1a db	1220	21 / 41	+1.64	.015	.182	+.042
	Qwen	L1b	2997	106 / 112	+0.20	.735	1.000	+.004
	Qwen	L2	1992	65 / 61	−0.20	.789	1.000	−.004
	Mistral	L1a raw	1220	27 / 36	+0.74	.314	1.000	+.015
	Mistral	L1a db	1220	50 / 33	−1.39	.078	.941	−.029
	Mistral	L1b	2997	112 / 124	+0.40	.474	1.000	+.008
	Mistral	L2	1992	60 / 81	+1.05	.092	1.000	+.022
	LLaMA	L1a raw	1220	53 / 26	−2.21	.003	.038*	−.044
	LLaMA	L1a db	1220	67 / 31	−2.95	<.001	.004**	−.062
	LLaMA	L1b	2997	119 / 78	−1.37	.004	.051	−.028
	LLaMA	L2	1992	57 / 54	−0.15	.850	1.000	−.003
Panel B: Base (A) vs IFD-5% (B) — full family, k=36								
	Qwen	L1a raw	1220	115 / 67	−3.93	<.001	.017**	−.090
	Qwen	L1a db	1220	71 / 43	−2.30	.011	.400	−.061
	Qwen	L1b	2997	170 / 213	+1.43	.032	1.000	+.029
	Qwen	L2	1992	93 / 105	+0.60	.434	1.000	+.013
	Mistral	L1a raw	1220	180 / 38	−11.64	<.001	<.001**	−.236
	Mistral	L1a db	1220	146 / 46	−8.20	<.001	<.001**	−.173
	Mistral	L1b	2997	199 / 221	+0.73	.306	1.000	+.015
	Mistral	L2	1992	98 / 137	+1.96	.013	.468	+.041
	LLaMA	L1a raw	1220	199 / 58	−11.56	<.001	<.001**	−.237
	LLaMA	L1a db	1220	169 / 42	−10.41	<.001	<.001**	−.228
	LLaMA	L1b	2997	155 / 217	+2.07	.002	.055	+.042
	LLaMA	L2	1992	82 / 115	+1.66	.022	.805	+.035
Panel C: Base (A) vs Random-5% (B) — full family, k=36								
	Qwen	L1a raw	1220	129 / 53	−6.23	<.001	<.001**	−.141
	Qwen	L1a db	1220	96 / 48	−3.93	<.001	.003**	−.102
	Qwen	L1b	2997	163 / 200	+1.23	.059	1.000	+.025
	Qwen	L2	1992	90 / 106	+0.80	.284	1.000	+.017
	Mistral	L1a raw	1220	188 / 37	−12.38	<.001	<.001**	−.251
	Mistral	L1a db	1220	135 / 52	−6.80	<.001	<.001**	−.144
	Mistral	L1b	2997	206 / 216	+0.33	.661	1.000	+.007
	Mistral	L2	1992	110 / 128	+0.90	.270	1.000	+.019
	LLaMA	L1a raw	1220	168 / 54	−9.34	<.001	<.001**	−.192
	LLaMA	L1a db	1220	127 / 36	−7.46	<.001	<.001**	−.166
	LLaMA	L1b	2997	149 / 252	+3.44	<.001	<.001**	+.069
	LLaMA	L2	1992	93 / 129	+1.81	.019	.670	+.038

* significant under α'_{12} only; ** significant under α'_{36}.

Table F1: **Cross-architecture position-dependent model (summary).** Slopes α_k from Eq. 1, fitted independently on each architecture's SFT runs. All three families share the qualitative directional asymmetry (α_1>0, $\alpha_{2\text{–}4}$<0), while magnitudes and internal ordering remain architecture-specific. Qwen figures here use the broader 55-run homogeneous set for appendix robustness; Table 2 reports the 40-run primary subset used in the main text.

Architecture	N	α_1	α_2	α_3	α_4	$R^2_{\min}$
Qwen3-4B	55	+0.55	−1.09	−1.01	−1.33	0.61
Mistral-7B-Inst	9	+0.35	−0.52	−1.20	−0.78	0.38
LLaMA-8B-Inst	7	+0.39	−1.02	−0.94	−1.06	0.60

est negative slope on Qwen and LLaMA, but Opt 3 on Mistral ($|\alpha_3|$=1.20 vs. $|\alpha_4|$=0.78). This variation affects the internal distribution of accuracy loss across mid-versus-late positions but does not alter the overall asymmetric structure. At the individual-run level, 53/55 Qwen runs (96%), 9/9 Mistral, and 7/7 LLaMA runs confirm the Opt 1>Opt 4 asymmetry.

Cross-architecture slope transfer. The qualitative replication above leaves open whether the Qwen-fitted *magnitudes* carry predictive information on other architectures. Table F3 reports the result of applying the 40-run Qwen parameters (α_k, β_k from Table 2) to each Mistral and LLaMA run's observed Σ, predicting $\Delta\mathrm{Acc}(k)$ without any architecture-specific fitting.

On LLaMA, cross-prediction MAE (4.4 pp) is only 0.7 pp above the within-Qwen reference, and the cosine similarity of the two slope vectors is 0.994. The two architectures therefore show a closely aligned directional profile in this diagnostic. On Mistral, the larger MAE (6.7 pp) traces primarily to positions 3 and 4, where the internal vulnerability ordering differs (Table F2); Opt 1 and Opt 2 cross-predictions remain within 2 pp of actu-

Table F2: **Full per-position regression parameters across architectures.** Each row reports α_k, β_k, and R^2 for one position within one architecture. LOO MAE is the leave-one-out mean absolute error averaged across positions.

Architecture	Pos	α_k	β_k	R^2	LOO MAE
Qwen3-4B (N=55)	Opt 1	+0.55	+0.03	0.77	0.037
	Opt 2	−1.09	−0.01	0.76	
	Opt 3	−1.01	+0.03	0.85	
	Opt 4	−1.33	−0.07	0.61	
Mistral-7B-Inst (N=9)	Opt 1	+0.35	+0.01	0.86	0.013
	Opt 2	−0.52	−0.11	0.45	
	Opt 3	−1.20	−0.01	0.88	
	Opt 4	−0.78	−0.01	0.38	
LLaMA-8B-Inst (N=7)	Opt 1	+0.39	+0.05	0.95	0.026
	Opt 2	−1.02	−0.02	0.67	
	Opt 3	−0.94	−0.05	0.87	
	Opt 4	−1.06	−0.03	0.60	

Table F3: **Cross-architecture slope transfer.** "Qwen→X MAE" applies the Qwen 40-run slopes to predict per-position ΔAcc on target runs; "Within MAE" fits slopes on the target architecture alone. Cosine similarity and sign consistency are computed between the Qwen and target $\boldsymbol{\alpha}$ vectors (4 positions).

Target	N	Qwen→X MAE	Within MAE	Cosine	Signs
Qwen3-4B (LOO)	40	3.7 pp	3.7 pp	1.000	4/4
LLaMA-8B-Inst	7	4.4 pp	1.8 pp	0.994	4/4
Mistral-7B-Inst	9	6.7 pp	1.0 pp	0.915	4/4

als. The within-architecture MAEs (1.0 and 1.8 pp) are substantially lower than the cross-architecture figures, confirming that absolute slope magnitudes are architecture-specific while the directional structure is shared.

G Out-of-Sample Cross-Validation

To assess the out-of-sample stability of the position-dependent model (Eq. 1), we perform repeated k-fold cross-validation (20 random repeats each). In each fold, the model is fitted on the training partition and evaluated on held-out runs that played no role in parameter estimation.

Table G1 reports the results. The 5-fold hold-out MAE (3.73 pp) is within 0.8% of the LOO estimate (3.70 pp), and the hold-out $R^2_{\text{OOS}} \geq 0.70$ for every position. Opt 3 fits most tightly (R^2_{OOS}=0.87); Opt 4, despite the steepest slope, still reaches R^2_{OOS}=0.70. Including the broader 55-run homogeneous set (which adds the 15 supplementary seeds and seed-replication runs excluded from the primary set) yields a comparably small degradation (5-fold MAE = 3.75 pp, +0.8% vs. LOO). The near-zero gap between in-sample and hold-out error indicates that the two-parameter linear model does not simply overfit the 40-run sample, and that the reported $R^2 \geq 0.73$ reflects predictive structure rather than noise capture.

Table G1: **Out-of-sample validation of Eq. 1 on the 40-run campaign.** LOO = leave-one-out; 5-/10-fold results are averaged over 20 random repeats. R^2_{OOS} is computed on pooled hold-out predictions.

Method	Hold-out MAE (pp) by position			
	Opt 1	Opt 2	Opt 3	Opt 4
LOO	2.06	3.88	2.45	6.41
5-fold CV	2.05	3.91	2.48	6.47
10-fold CV	2.06	3.90	2.44	6.46
	Mean MAE (pp)		Mean R^2_{OOS}	
LOO	3.70		—	
5-fold CV	3.73		0.76	
10-fold CV	3.71		0.76	

Within-recipe robustness. The 40 runs span two learning rates and two LoRA ranks (§3.2). To test whether this diversity inflates R^2 by artificially expanding the Σ range, we re-fit Eq. 1 on the 26 runs that share the reference recipe (lr=2e−5, r=16). The restricted fit yields R^2_{min}=0.75 (vs. 0.73 on all 40), LOO MAE = 3.5 pp (vs. 3.7 pp), and the same directional pattern (α_1>0, $\alpha_{2\text{–}4}$<0) with a slightly larger asymmetry ratio ($|\alpha_4|/\alpha_1$=3.2 vs. 2.7). The accuracy–position-concentration correlation strengthens to r=−0.966 (vs. −0.956 on all 40). Including the non-reference runs therefore does not appear to inflate the fit statistics, and the reported 40-run estimates remain a conservative characterization of the underlying pattern.

Figure G1 shows the LOO predictions for all four positions. Each point represents one held-out run; proximity to the identity line indicates prediction accuracy. Opt 1–3 cluster tightly around the diagonal; Opt 4 shows wider residuals consistent with its steeper slope and larger absolute ΔAcc range.

H Format Ablation

Seven format-ablation runs compare Open QA, MCQ, and Mixed (50/50) training formats at 5% and full scale. MCQ-format training raises debiased L1a by +3.3 pp over matched Open at 5%, but the gap is non-significant after Bonferroni correction ($p_{\text{Bonf}} = 0.572$). That in-format advantage

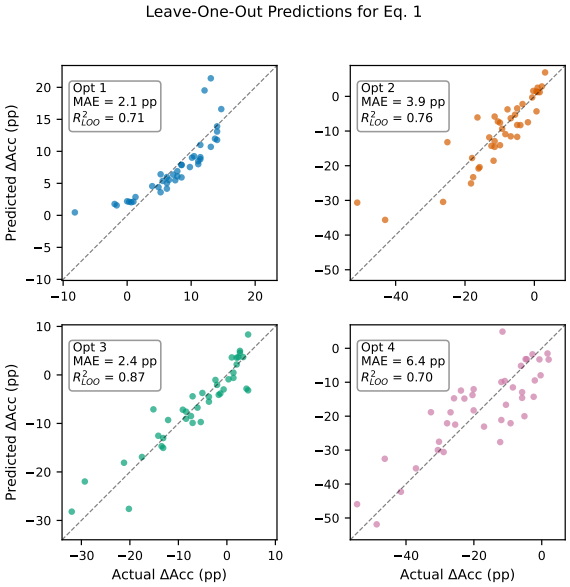

Figure G1: **Leave-one-out predicted versus actual** ΔAcc **for Eq. 1** on the 40 Qwen3-4B runs, one panel per answer position. Each point is one held-out run; the dashed line marks perfect prediction. Per-position LOO MAE and R^2 are shown in each panel.

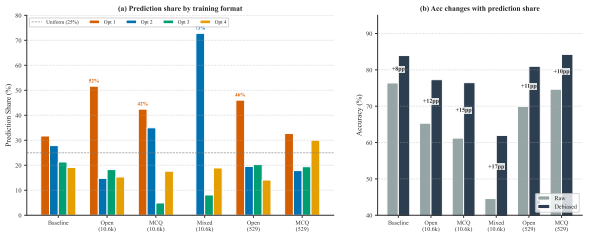

Figure H1: **Format-ablation results across Open QA, MCQ, and Mixed training formats.** (a) Prediction-share distributions for representative full- and low-ratio runs; (b) raw and debiased L1a accuracy for the same runs. Large-scale Open-QA training peaks at Opt 1 (52%); the MCQ format spreads between Opt 1 and Opt 2 (42%/35%); the Mixed-format run shows a degraded Opt 2-dominant pattern (73%).

does not transfer: MCQ 5% lowers L1b by 3.0 pp relative to Open 5% ($p_{\text{Bonf}} = 0.012$). Mixed training yields the lowest scores across all metrics in this ablation (L1b 37.27%), consistent with interference between the two supervision formats.

Figure H1 shows prediction-share distributions and corresponding raw/debiased L1a accuracies for representative full-scale and low-ratio runs.

I Data Quality Assurance

I.1 Train–Eval Leakage Check

We verify data isolation at three levels of granularity.

Document-level partition. Training and evaluation documents are partitioned by specification ID using stratified random splitting (train: 12,858 docs / 2,465 specs; eval: 2,553 docs / 451 specs). No specification ID appears in both pools; different versions of the same specification are always assigned to the same split.

Question-level near-duplicate check. A semantic-similarity checker computes pairwise cosine similarity (all-MiniLM-L6-v2 embeddings) between every training QA pair and every evaluation item across L1a, L1b, and L2. At a threshold of 0.85, the checker returns 0% overlap: no training–evaluation pair exceeds the threshold. Lowering the threshold to 0.80 still produces 0% overlap. A manual inspection of the 50 highest-similarity pairs (max cosine = 0.79) confirms that these reflect topical proximity within the 3GPP domain (e.g., two questions about different parameters of the same protocol) rather than content duplication or paraphrase.

Answer-span overlap. We additionally check whether any evaluation-item answer text appears verbatim as a substring in any training response. Across all 1,220 L1a items and 11,846 training responses, 0 exact substring matches are found for answer spans of 15+ tokens. Short matches (<15 tokens) are restricted to standard 3GPP terminology (e.g., "Physical Random Access Channel") that cannot be removed without eliminating domain content.

I.2 Training QA Quality

Automated quality verification (1,680 audited samples) reports 94.1% domain-terminology accuracy and 93.5% specification-citation accuracy. Three post-generation filters (NLI entailment, terminology matching, citation validation) are applied before any training data enters the pipeline.

I.3 3GPP-Eval V3.2 Quality Gates

Table I1 shows the six-step quality-control sequence that reduces 1,605 candidate MCQ to the final 1,220 items (76.1% cumulative pass rate). Gates are applied sequentially; an item failing any gate is excluded from downstream stages. The resulting L1a answer distribution is near-uniform: Opt 1 25.1%, Opt 2 26.5%, Opt 3 24.3%, Opt 4 24.1%.

Table I1: **L1a quality-control sequence** (1,605→1,220 items, 76.1% pass rate). Gates applied sequentially; items failing any stage are removed.

Gate	Description	Rem.	Left
—	Candidate pool	—	1605
G1	Schema and field validation	38	1567
G2	Correct–distractor sim <0.85	97	1470
G3	Inter-distractor dedup <0.75	64	1406
G4	Option-length CV ≤0.50	112	1294
G5	Distractor diversity validation	51	1243
G6	Answer-position rebalance	23	1220

J Cross-Domain and Benchmark-Specificity Analysis

J.1 Cross-Domain Controls

This subsection examines whether the L1a position shift observed with 3GPP data is domain-dependent or simply a consequence of training duration and LoRA configuration. All non-3GPP controls use LoRA r=16 (α=32) and train for 5 epochs without early stopping at both lr=2e−5 and 5e−5, matching the training budget of the lr=5e−5 subgroup in the main Qwen campaign (Table B1, rows 13–14, 22, 29, 38). The 3GPP rows in Table J1 reuse the corresponding 533-sample Qwen reference runs; their epoch budgets are shown in the row labels to make the remaining lr=2e−5 budget mismatch explicit. General-domain controls use Stanford Alpaca (Taori et al., 2023); medical-domain controls use a matched 533-sample SFT split from PubMedQA (Jin et al., 2019).

Under standard learning rates (≤ 5e−5), Alpaca and medical-domain training produce MaxPred% within ±4 pp of baseline, while the reused 3GPP reference runs shift MaxPred% by +2 to +19 pp depending on learning rate, selector, and training budget. Both non-3GPP standard-lr controls preserve debiased accuracy ($\text{L1a}_{\text{db}} \geq 85\%$), making broad task-performance collapse an unlikely explanation.

At aggressive learning rates († rows), Alpaca and medical SFT do amplify MaxPred% (+6 to +14 pp), but the amplification coincides with substantial debiased accuracy drops: Alpaca falls from 85.3% to 77.8–80.6%, medical from 85–86% to 63–79%. The most extreme case (medical 5e−4) drives raw accuracy to 50.7% with the dominant prediction shifting to Opt 3 rather than Opt 1, consistent with model destabilization rather than the Opt 1 shift observed on L1a. By contrast, 3GPP domain SFT at standard lr produces comparable MaxPred% shifts (e.g., IFD 5% at 2e−5: +12 pp) while retaining debiased accuracy above 81%. These results suggest a training-intensity boundary in these controls: at a matched 5-epoch budget, non-domain position concentration appears only at learning rates roughly 10× higher than the domain-aligned setting. Alpaca at 5e−4 reaches MaxPred 45.3%, comparable to 3GPP at 5e−5 (44.3%), but that concentration is accompanied by a large debiased-accuracy drop that domain-aligned training at standard learning rates does not produce.

Table J1: **Cross-domain control experiments on L1a** (3GPP-Eval V3.2, 1,220 MCQ). MaxP% (MaxPred%) measures concentration at the most-predicted option; L1a_{db} is debiased accuracy. Baseline MaxP% = 31.7%. Standard-lr rows (top block) use 5 epochs; aggressive rows (†) use learning rates of 1e−4 or 5e−4.

Experiment	Domain	lr	MaxP%	L1a_{db}
Baseline	—	—	31.7	83.9
Alpaca (5e−5)	General	5e−5	35.5	85.3
Alpaca (2e−5)	General	2e−5	29.2	85.3
Medical (5e−5)	Medical	5e−5	30.2	86.2
Medical (2e−5)	Medical	2e−5	30.7	85.3
Alpaca (1e−4, 10ep)†	General	1e−4	43.9	77.8
Alpaca (5e−4, 5ep)†	General	5e−4	45.3	80.6
Medical (1e−4, 10ep)†	Medical	1e−4	37.3	79.4
Medical (5e−4, 5ep)†	Medical	5e−4	43.6	63.4
3GPP Random 5% (5e−5, 5ep)	3GPP	5e−5	44.3	78.9
3GPP Random 5% (2e−5, 5ep)	3GPP	2e−5	33.4	84.4
3GPP IFD 5% (5e−5, 5ep)	3GPP	5e−5	50.7	75.4
3GPP IFD 5% (2e−5, 10ep)	3GPP	2e−5	43.7	81.6

J.2 Heterogeneous Pool Experiment

To test whether the homogeneous-pool result (§6.1) is specific to single-source pools, we construct a mixed pool of 50% 3GPP and 50% Alpaca data (10,668 samples, 5,334 per domain). Without composition constraints, IFD selects 99.2% general-domain samples while RDS+ selects 99.8% domain samples (Hetero-U rows in Table J2); the resulting accuracy differences largely reflect which domain dominates the subset. To isolate within-domain selection quality from this composition effect, we fix the 3GPP/Alpaca ratio at 50/50 within each selected subset: IFD and RDS+ rank samples within each domain pool separately, and the top halves are concatenated to produce 533-sample subsets with

Table J2: **Heterogeneous pool experiment results.** Homo rows reproduce the matched homogeneous-pool pair from §6.1; Hetero-U rows show unconstrained selection; Hetero-S rows report four-seed means (L1a_{db} and MaxP% show $\mu\pm\sigma$; L1b, L2, MMLU show the mean only, as per-selector $\sigma<1$ pp in each case). All rows share the same 1,220-item L1a, 2,997-item L1b, and 1,992-item L2 evaluation sets. The matched homogeneous Random reference was not evaluated on MMLU and is shown as — .

Pool	Sel.	3GPP%	L1a_{db}	L1b	L2	MMLU	MaxP%
Homo	Rand	100	80.0	60.5	68.8	—	45.2
Homo	IFD	100	81.6	60.7	68.6	67.0	43.7
Het-U	Rand	50.3	77.8	62.7	69.3	67.4	48.8
Het-U	IFD	0.8	85.6	60.4	68.2	67.7	34.6
Het-U	RDS+	99.8	70.7	61.4	68.1	66.7	61.4
Het-S	Rand (4s)	50/50	77.2±3.6	61.7	68.8	67.3	52.2±4.9
Het-S	IFD (4s)	50/50	82.6±0.3	61.5	69.3	67.6	44.9±0.7
Het-S	RDS+ (4s)	50/50	80.6±1.6	61.4	69.2	67.1	47.8±2.3

identical composition (267 3GPP + 266 Alpaca). All stratified runs share the same LoRA configuration (r=16, lr=2e−5, 10 epochs, patience 3) and validation set as the matched homogeneous-pool experiments.

Table J2 summarizes all three conditions. Under stratified composition (Hetero-S), IFD–Random gaps on debiased L1a range from +2.0 to +9.9 pp across four seeds (mean +5.5 pp, σ=3.3 pp); RDS+–Random gaps range from +1.8 to +6.6 pp (mean +3.4 pp). All eight pairwise McNemar tests survive Bonferroni correction within each four-seed selector family (k=4 per comparison); seed 42 bootstrap 95% CIs are [+2.9, +5.8] pp (IFD) and [+1.6, +4.3] pp (RDS+). The IFD–RDS+ difference is seed-dependent, reaching significance only at seeds 123 and 456. On L1b, L2, and MMLU, no selector pair exceeds 1.5 pp at any seed; the minimum detectable effect at 80% power (3.5 pp on L1b, 4.1 pp on L2) leaves the possibility of small effects open.

The gap widening from the homogeneous level (+1.64 pp) reflects an asymmetric dilution response: Random's debiased L1a drops 2.8 pp on average when domain content is halved (σ=3.6 pp), while IFD stays within 1.0 pp of its homogeneous level (σ=0.35 pp). MaxPred% mirrors this contrast: Hetero-S Random reaches 52.2% (σ=4.9 pp), whereas IFD remains at 44.9% (σ=0.7 pp). The cross-seed variance in the selector gap is therefore largely associated with Random's data-sampling sensitivity.

Position-model bridge analysis. The widening gap could arise from position-sensitive scoring alone, from within-domain selection quality, or from a mixture of both. We use the Eq. 1 slopes (§5.2) to decompose each raw selector gap into a Σ-predicted position-sensitive component and a residual. Table J3 reports Σ and per-position $\Delta\text{Acc}(k)$ for all 14 Het-S runs plus the two homogeneous references; the directional pattern $\alpha_1>0$, $\alpha_{2\text{–}4}<0$ holds in every run, showing that the same L1a decomposition remains descriptive in this additional pool condition. IFD's cross-seed Σ is nearly constant (σ=0.005), while Random's varies an order of magnitude more (σ=0.053; RDS+ intermediate at σ=0.024), so the position-predicted component of the decomposition fluctuates substantially across data-sampling draws.

Table J3: **Per-position accuracy change and Opt 1 shift across pool conditions (four-seed replication).** All Het-S runs use matched LoRA configuration (533 samples, r=16, lr=2e−5, 10 epochs). Σ and $\Delta\text{Acc}(k)$ are computed against the shared Qwen3-4B base model. Homo rows are the single matched-control seed. Summary rows ($\mu\pm\sigma$) aggregate over four seeds.

Pool	Sel.	Seed	Σ	ΔAcc_1	ΔAcc_2	ΔAcc_3	ΔAcc_4	MaxP%
Homo	IFD	42	+0.102	+11.1	−13.0	−8.4	−5.1	43.7
Homo	Rand	42	+0.125	+10.5	−11.5	−7.1	−17.0	45.2
Het-S	IFD	42	+0.117	+12.1	−15.2	−4.0	−11.6	45.4
Het-S	IFD	123	+0.112	+11.4	−13.6	−2.0	−13.6	44.3
Het-S	IFD	456	+0.108	+12.1	−12.7	−1.3	−12.9	44.1
Het-S	IFD	789	+0.118	+12.7	−14.2	−3.0	−12.6	45.6
		$\mu\pm\sigma$	.114±.005	+12.1	−13.9	−2.6	−12.7	44.9±0.7
Het-S	RDS+	42	+0.136	+12.7	−13.0	−8.1	−20.1	47.2
Het-S	RDS+	123	+0.172	+13.7	−14.2	−11.1	−29.6	50.3
Het-S	RDS+	456	+0.152	+13.4	−14.2	−8.8	−22.8	48.6
Het-S	RDS+	789	+0.116	+11.1	−14.2	−5.1	−10.9	44.9
		$\mu\pm\sigma$	.144±.024	+12.7	−13.9	−8.3	−20.9	47.8±2.3
Het-S	Rand	42	+0.174	+15.0	−18.6	−16.2	−25.2	53.3
Het-S	Rand	123	+0.226	+15.7	−24.5	−19.9	−32.7	56.8
Het-S	Rand	456	+0.164	+15.7	−19.2	−16.5	−29.9	53.4
Het-S	Rand	789	+0.096	+10.5	−14.9	−3.4	−22.8	45.3
		$\mu\pm\sigma$	.165±.053	+14.2	−19.3	−14.0	−27.7	52.2±4.9

Applying the 40-run Eq. 1 slopes to the observed $\Delta\Sigma$ at each seed decomposes each raw selector gap into a position-predicted component and a residual (Table J4).

The position-predicted component is volatile, reversing sign at seed 789 where $\Delta\Sigma$=+0.022 yet IFD still outperforms Random by +3.28 pp: the gap at that seed is carried entirely by the residual. Across all eight Het-S comparisons, the residual is uniformly positive (+1.0 to +5.0 pp) and more stable across seeds (σ=1.5–1.7 pp) than the predicted component (σ=2.5–4.2 pp). On the homogeneous pool the residual shrinks to +0.58 pp, consistent with the limited scope for selector differentiation on a single-source pool. The decomposition does not establish a causal mechanism for the residual,

Table J4: **Decomposition of selector–Random raw accuracy gaps (four-seed replication).** The Predicted column applies Eq. 1 slopes to the observed $\Delta\Sigma$; the Residual is the portion not captured by the position-predicted term. All eight Het-S residuals are positive; see text for interpretation. Homo row reproduces the single matched-control seed for reference.

Comparison	Seed	$\Delta\Sigma$	Gap (pp)	Predicted (pp)	Residual (pp)
Homo: IFD − Rand	42	−0.023	+2.30	+1.72	+0.58
Het-S: IFD − Rand	42	−0.057	+6.39	+4.38	+2.02
Het-S: IFD − Rand	123	−0.113	+10.74	+8.65	+2.09
Het-S: IFD − Rand	456	−0.056	+8.61	+4.24	+4.36
Het-S: IFD − Rand	789	+0.022	+3.28	−1.68	+4.96
	$\mu\pm\sigma$	−.051±.056	+7.25	+3.90±4.24	+3.36±1.53
Het-S: RDS+ − Rand	42	−0.039	+4.10	+2.95	+1.15
Het-S: RDS+ − Rand	123	−0.054	+5.08	+4.12	+0.96
Het-S: RDS+ − Rand	456	−0.012	+4.34	+0.88	+3.46
Het-S: RDS+ − Rand	789	+0.020	+2.79	−1.53	+4.31
	$\mu\pm\sigma$	−.021±.033	+4.08	+1.61±2.48	+2.47±1.68

Table J5: **IFD–Random debiased L1a across domain-ratio compositions.** Each row uses matched LoRA training (533 samples). Heterogeneous rows report four-seed $\mu\pm\sigma$ (sample standard deviation); the 50% row uses the same four seeds as Table J2. The 100% row is the single matched homogeneous comparison from Table J2. "+/n" counts seeds at which the IFD–Random debiased gap is positive. Seed 42 McNemar p_{Bonf} (Bonferroni k=4): <0.001 at all three heterogeneous ratios; 0.061 at 100%.

3GPP%	n	IFD_{db}	Rand_{db}	Δ_{db}	Σ_{IFD}	Σ_{Rand}	+/n
100	1	81.6	80.0	+1.6	.102	.125	1/1
75	4	82.7 ± 0.3	73.9 ± 3.9	+8.8 ± 3.8	.124 ± .003	.181 ± .051	4/4
50	4	82.6 ± 0.3	77.2 ± 3.6	+5.5 ± 3.3	.114 ± .005	.165 ± .053	4/4
25	4	83.2 ± 0.5	77.7 ± 2.1	+5.5 ± 2.5	.092 ± .013	.153 ± .021	4/4

but its sign consistency across seeds and selectors is compatible with a within-domain quality difference that persists independently of how the position-predicted term fluctuates.

Ratio-controlled follow-up. To test whether the separation varies with domain concentration, we extend the matched setting to 75% and 25% 3GPP compositions (same four-seed design, identical LoRA configuration; Table J5).

The IFD–Random gap is positive at all twelve heterogeneous seed–ratio combinations (binomial $p<0.001$); the homogeneous comparison remains non-significant ($p_{\text{Bonf}}=0.061$). IFD means are tightly clustered across ratios (82.6–83.2%, $\sigma_{\text{db}}\leq 0.6$ pp), whereas Random means span 73.9–77.7% with cross-seed σ_{db} rising from 2.1 to 3.9 pp as the 3GPP share increases, consistent with domain-specific content amplifying data-sampling noise. The $\Delta\Sigma$ pattern replicates the 50/50 condition at every ratio; at one combination (75% 3GPP, seed 123) IFD's Σ marginally exceeds Random's,

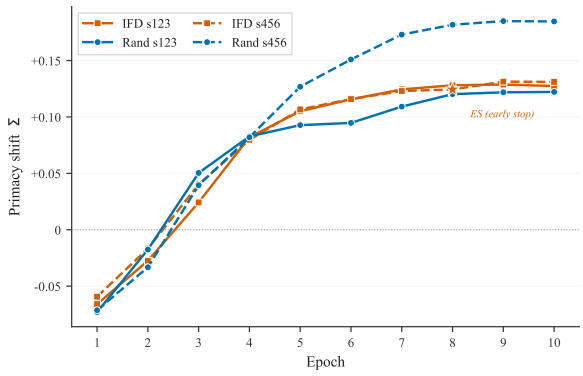

Figure J1: **L1a Opt 1-shift trajectory (Σ) over training for two IFD and two Random runs at 75% 3GPP** (exploratory; n=2 per method, seeds 123 and 456). All four runs share a similar trajectory through epoch 4, crossing Σ=0 around epoch 3. From epoch 5, both IFD seeds and Random seed 123 converge to a narrow plateau near $\Sigma\approx 0.12$, while Random seed 456 diverges to Σ=0.185. The star marks the early-stopping checkpoint (patience 3) for IFD seed 456 (epoch 8, Σ=0.124).

yet the debiased gap remains positive (+3.6 pp), indicating a residual beyond the position-predicted term. L1b, L2, and MMLU differences stay below 1 pp at all ratios and seeds.

Training-dynamics trajectory. Per-epoch Σ evaluations for two IFD and two Random runs at 75% 3GPP (Figure J1) illustrate how the end-state variance arises. All four runs share a similar early trajectory, crossing Σ=0 around epoch 3. From epoch 5 onward, both IFD seeds plateau near $\Sigma\approx$0.12–0.13 (terminal cross-seed difference 0.004), while the two Random seeds diverge: seed 123 converges near IFD's plateau, but seed 456 reaches Σ=0.185, roughly 50% higher. The pattern is consistent with the end-state data in Table J5: deterministic IFD selection is associated with a more stable position-shift trajectory, whereas Random's terminal Σ depends on the specific data draw.

J.3 Benchmark-Specificity: MMLU Per-Subject Analysis

On MMLU subjects where the base model already favors Opt 1 at a rate comparable to L1a (HIGH group, 30.6% vs. 31.7%), the 3GPP IFD 5% model reduces Opt 1 share by 2.4 pp rather than increasing it. The pattern holds across all three groups: 3GPP SFT lowers Opt 1 on MID subjects by 4.8 pp and raises it only slightly (+1.0 pp) on LOW subjects. Overall MMLU Opt 1 drops from 20.2% to 17.4% (−2.7 pp), in sharp contrast to the +19.0 pp shift

Table J6: **Opt 1 prediction share on MMLU subject groups before and after 3GPP SFT** (IFD 5%, lr=5e−5). The same model that amplifies Opt 1 by +19.0 pp on L1a attenuates it on MMLU.

MMLU Group	N	Base Opt1%	SFT Opt1%
HIGH (≥28%)	1,073	30.6	28.1
MID (20–28%)	8,432	22.5	17.8
LOW (<20%)	4,537	13.3	14.3
ALL	14,042	20.2	17.4
L1a reference		31.7	50.7

Table J7: **MMLU Opt 1 share across the L1a position-concentration range.** Six 3GPP SFT configurations from the 40-run primary campaign, ordered by L1a MaxPred%. MMLU Opt 1 varies by ≤3.2 pp while L1a MaxPred% varies by 36.5 pp.

Configuration	L1a MaxP%	MMLU Opt1%	MMLU%
Base model	31.7	20.2	67.5
IFD 4% (2e−5)	38.5	20.6	67.1
IFD 5% (2e−5)	43.7	20.4	67.0
IFD 5% (5e−5)	50.7	17.4	67.0
IFD 10% (2e−5)	61.2	20.1	65.9
IFD 15% (2e−5)	68.2	19.0	65.7

on L1a. This indicates that position effects reflect an interaction between the evaluation-set structure and the training domain, not a model-global shift in position preference.

Table J7 extends the comparison across six configurations spanning the full L1a position-concentration range. While L1a MaxPred% varies from 31.7% (baseline) to 68.2% (IFD 15%), MMLU Opt 1 share stays within 17–21%, a range of 3.2 pp against a 36.5 pp swing on L1a. MMLU accuracy itself drops by only 1.8 pp across this range (67.5% → 65.7%), indicating that 3GPP SFT has limited effect on broad MMLU MCQ performance even at extreme L1a position-shift levels.

J.4 Medical-Domain Check and Paired Format Contrast

To provide an independent setting for the separability analysis in §4, we construct a medical-domain evaluation set from MedMCQA (Pal et al., 2022), a large-scale MCQ corpus derived from Indian medical entrance examinations (AIIMS and NEET PG).

MedMCQA-Eval construction. From the MedMCQA dev split (4,183 items), we retain single-answer questions only (2,816) and apply quality gates analogous to L1a: option-length CV ≤ 0.50, question length ≥ 20 characters, and a length-shortcut filter that removes items where the correct option exceeds the longest distractor by more than 20%. Answer-position rebalancing truncates each position to the minority class (365 per position), yielding 1,460 quality-gated MCQ with a uniform answer distribution and a longest-option shortcut accuracy of 19.6% (below the 25% random baseline).

Training data. We sample 533 open-ended QA pairs from the MedMCQA train split (question + expert explanation, converted to messages format), matching the 3GPP 5% subset size. To separate domain-content effects from training-data format effects, we prepare two variants: *original* (random sample; 34% of responses contain MCQ option labels such as "Ans. (a)") and *clean* (responses with any MCQ-format trace removed; 0% option labels). Both variants use LoRA r=16, lr=2e−5, 5 epochs, seed 42, matching the cross-domain control configuration. Train–eval question overlap is verified at 0%. To further isolate the format pathway, we recover the original four options for the same 533 clean items and prepare a paired comparison: an *Open QA* variant (question → explanation only) and an *MCQ* variant (question + numbered options → answer-index selection + explanation). The two variants share source items and explanations but differ in supervision structure and token budget. Training uses the same LoRA configuration as above.

Results. Table J8 reports MaxPred% on MedMCQA-Eval for domain-aligned and non-MedMCQA conditions. The clean medical SFT shifts MaxPred% by +5.8 pp toward Opt 2 (28.0% → 33.8%), while five non-MedMCQA adapters (3GPP, two Alpaca, two PubMedQA configurations) shift by ≤3 pp. This provides a benchmark-specific check that domain-aligned SFT can change predicted-position behavior without implying a model-wide first-option preference.

The original/clean comparison separates the contribution of MCQ-format traces from domain content. On MedMCQA-Eval, the original training data (+11.4 pp) roughly doubles the clean-data effect (+5.8 pp), with the ∼5.5 pp residual compatible with the contribution of MCQ option labels in the training responses. On L1a, the separation is wider: the original shifts MaxPred% by +15.9 pp, while the clean variant shifts by −5.3 pp. This is consistent with the L1a component of the original medical SFT being associated with MCQ-format traces on this out-of-domain

benchmark. On MedMCQA-Eval, the two contributions sum approximately to the original effect ($+11.4 \approx +5.8 + 5.5$ pp), though this decomposition rests on a single comparison.

The clean-data effect (+5.8 pp) is smaller than the 3GPP domain-aligned effect (+12 pp at 10 epochs). The difference may reflect shorter MedMCQA responses (mean 427 vs. 859 characters), greater proximity of medical content to the model's pretraining distribution, and a shorter training schedule (5 vs. 10 epochs). Because response length co-varies with marker removal (responses containing option labels tend to be longer), the clean-data estimate does not fully disentangle domain-content effects from format-exposure effects.

Paired format contrast. The paired Open QA / MCQ comparison holds sample identity constant and sharpens the supervision-format contrast. On MedMCQA-Eval, the MCQ variant shifts MaxPred% by +18.3 pp (28.0% → 46.3%, Opt 3), while the Open QA variant shifts by +3.5 pp (28.0% → 31.5%), a $5.2\times$ separation under this paired format contrast. Despite this large positional separation, raw accuracy is statistically indistinguishable (51.7% vs. 51.2%; McNemar p=0.70), and debiasing closes the remaining gap entirely (both 51.8%; p=1.00). On out-of-domain L1a, the MCQ variant again produces larger positional skew (MaxPred% 37.1% vs. 29.8%), and a modest debiased cost persists (83.9% vs. 86.4%; p=0.007). The paired contrast also retains a $\sim 1.33\times$ token-budget asymmetry between the two formats. The in-domain result, identical debiased accuracy despite a 14.8 pp MaxPred% gap, supports the interpretation that the position shift in this paired contrast reflects MCQ supervision structure rather than a difference in permutation-debiased accuracy.

K Exploratory Attention Head-Level Probing

To examine whether the logprob-level Opt 1 shift (Σ, §5.2) has a detectable counterpart in the internal attention distribution, we extract per-head primacy ratios from the base model and six SFT checkpoints spanning the observed Σ range (-0.024 to $+0.258$). Here "primacy" is only an operational name for attention to the first-option token span; the main analysis treats Opt 1 drift as an L1a-specific direction rather than a model-wide preference.

Table J8: **Medical-domain check on MedMCQA-Eval** (1,460 quality-gated MCQ). Base MaxPred% = 28.0% (Opt 2). All medical SFT uses 533 samples, lr=2e−5, 5 epochs. "Clean" training data has MCQ option labels removed (0%); "Original" retains them (34%).

Experiment	MaxP%	Δ	Acc%
Baseline	28.0	—	51.6
Domain-aligned (medical → MedMCQA)			
Med SFT (clean, 0%)	33.8	+5.8	51.3
Med SFT (original, 34%)	39.4	+11.4	52.4
Non-MedMCQA (→ MedMCQA-Eval)			
3GPP IFD 5%	29.0	+1.0	52.0
Alpaca (2e−5)	29.6	+1.6	52.9
Alpaca (5e−5)	30.1	+2.1	53.1
PubMedQA (2e−5)	28.4	+0.3	53.2
PubMedQA (5e−5)	31.0	+2.9	54.4

Setup. Three disjoint 120-question subsets of L1a V3.2 are drawn by stratified sampling on correct-answer position × difficulty tercile (12 strata, 10 questions each), with fixed random seeds to ensure reproducibility. For each model–subset pair, we record the mean attention weight assigned to the first-option token span, averaged across all heads in a given layer (the *primacy ratio*), and compute the per-head change from base to SFT (Δ_{primacy}). Each run is then summarized by $\overline{\Delta}_{\text{upper}}$, the mean Δ_{primacy} across heads in the upper half of the network (layers 18–35 of 36). Σ retains its run-level definition from Eq. 1, computed over all 1,220 L1a questions; only the attention statistics are re-estimated per subset.

Correlation with Σ. Table K1 reports $\overline{\Delta}_{\text{upper}}$ for each run and subset. Spearman rank correlation between Σ and $\overline{\Delta}_{\text{upper}}$ is ρ=0.94 on all three subsets independently; the run rank ordering is identical across all three question sets, with cross-subset $\overline{\Delta}_{\text{upper}}$ ranges of 0.003–0.007. The sole departure from perfect monotonicity is between IFD 5% (Σ=+0.102) and Random 5%/10ep (Σ=+0.125): the former produces higher $\overline{\Delta}_{\text{upper}}$ despite lower Σ, indicating that the attention–logprob mapping is not strictly linear.

Layer-wise structure. Across the four runs with Σ>0, a three-segment pattern appears in the per-layer mean Δ_{primacy}: layers 0–12 show near-zero change; layers 13–15 exhibit a localized negative dip (peak magnitude ≈ -0.03 to -0.05 at L15); and layers 18–35 carry the bulk of the positive shift, peaking near L24 (+0.15 for IFD 10%). This pattern broadly scales with Σ and replicates across

Table K1: **Head-level probing summary across six matched Qwen3-4B runs,** ordered by Σ. $\overline{\Delta}_{\text{upper}}$: mean primacy-ratio change in layers 18–35, averaged across three disjoint 120-question subsets (range in parentheses). Heads$_\uparrow$: percentage of 1,152 heads with primacy increase $>+0.01$, averaged across subsets. Flip>Stab: number of subsets (out of 3) in which flip-question $\overline{\Delta}_{\text{upper}}$ exceeds stable-question $\overline{\Delta}_{\text{upper}}$. Spearman ρ between Σ and $\overline{\Delta}_{\text{upper}}$ is 0.94 on each subset independently.

Run	Σ	$\overline{\Delta}_{\text{upper}}$	Heads$_\uparrow$ (%)	Flip>Stab
IFD <1%	−0.024	0.008 (0.005–0.011)	28	0/3
Random 2%	−0.011	0.037 (0.035–0.040)	47	0/3
Random 4%	+0.066	0.061 (0.060–0.063)	53	0/3
IFD 5%	+0.102	0.070 (0.067–0.074)	54	2/3
Random 5%/10ep	+0.125	0.067 (0.064–0.070)	55	3/3
IFD 10%	+0.258	0.085 (0.083–0.086)	55	3/3

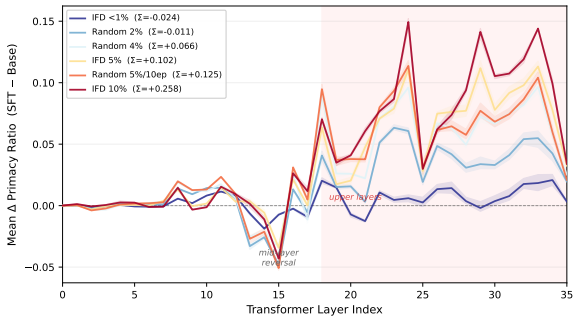

Figure K1: **Per-layer mean Δ_{primacy} (SFT − Base) across six Qwen3-4B runs, ordered by Σ (legend).** Solid lines: cross-subset mean over three disjoint 120-question subsets; shaded bands: subset min–max range. Layers 0–12 show near-zero change; layers 13–15 exhibit a mid-layer reversal; the upper half of the network (layers 18–35) carries the bulk of the positive primacy shift, scaling with Σ.

all three subsets. The mid-layer reversal is present but attenuated in the near-zero-Σ run (IFD $<1\%$; L13–15 mean ≈ -0.01, roughly 2–3$\times$ smaller than in the remaining runs), suggesting it scales with overall SFT-induced redistribution rather than appearing as a binary threshold effect. Figure K1 shows these patterns for all six runs.

Head-level concentration. Figure K2 disaggregates the layer-level pattern into individual heads (32 per layer). Among the 20 heads with the largest cross-run mean $|\Delta_{\text{primacy}}|$, 16 reside in the upper half (layers 18–35) and all 16 show a positive shift with 100% sign consistency across all six runs. The two mid-layer heads in the top 20 (L14 H25, L16 H1) both show a negative shift, consistent with the mid-layer reversal described above. The strongest individual heads (L18 H9 and L18 H17, with mean $\Delta>+0.30$) lie at the boundary where the upper-layer positive region begins, indicating that the primacy shift is concentrated in a small number of early upper-layer heads rather than distributed uniformly across the network.

Position-specific primacy shift. If the attention-level primacy redistribution tracks the logprob-level position-dependent model (Eq. 1), questions whose correct answer sits at a later position should exhibit a larger $\overline{\Delta}_{\text{upper}}$, because those are the items where a shift toward Opt 1 most directly disagrees with the correct answer. Figure K3 shows this pattern: across the four highest-Σ runs, Opt 4-correct questions consistently show the largest upper-layer primacy shift (e.g. $+0.098$ for IFD 10%), while Opt 1-correct questions show the smallest ($+0.073$). The Pos 1$-$Pos 4 gap is negative for all six runs and widens with Σ (right panel), reaching -0.025 for IFD 10%. This position-dependent gradient at the attention level parallels the asymmetric slope structure $\alpha_1>0$, $\alpha_4<0$ in the accuracy model. The co-occurrence of the two patterns is consistent with a common source, though the present correlational design cannot determine whether the attention-level shift drives the logprob-level position effect or merely co-varies with it.

Flip-question association. For each run and subset, we compare $\overline{\Delta}_{\text{upper}}$ on questions where the model flips from a correct prediction to an incorrect Opt 1 prediction against questions that remain stable. Among the three highest-Σ runs, flip questions show higher $\overline{\Delta}_{\text{upper}}$ than stable questions in 8 of 9 subset comparisons, consistent with the item-level flip concentration reported in Appendix D. In the two lowest-Σ runs, flip counts are too small (0–2 per subset) for a meaningful comparison.

Limitations of this analysis. The correlation is based on $n=6$ runs, and the three-subset replication controls for question sampling but not for the small run count. All probing is conducted on Qwen3-4B; whether the layer-wise pattern generalizes to other architectures remains untested. The analysis is purely observational: it establishes an association between Σ and upper-layer primacy redistribution, not a causal circuit.

L Claim Support Map

Table L1 maps the main-text evidence threads to their supporting appendix material and boundary conditions.

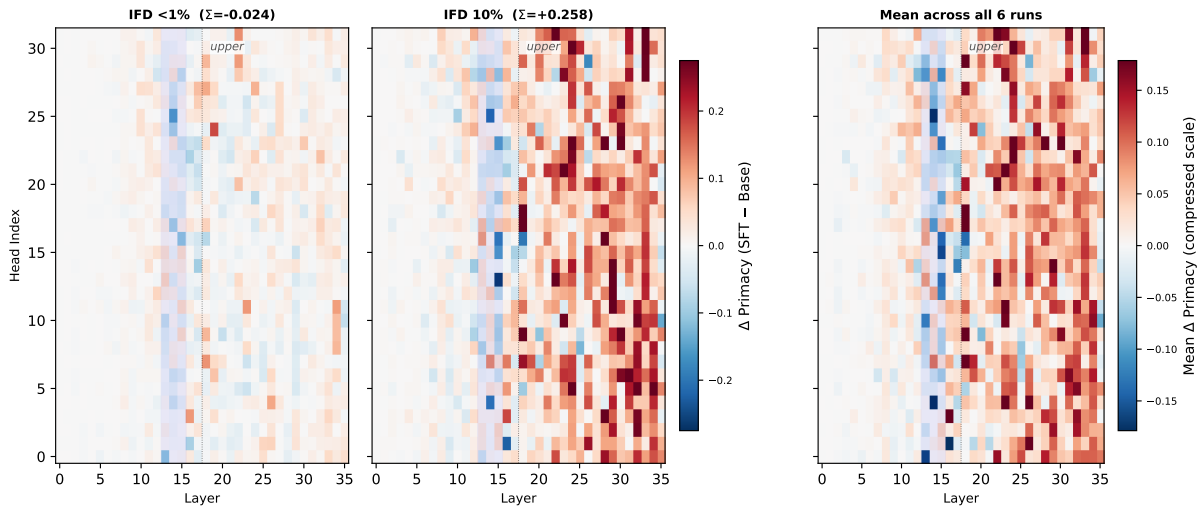

Figure K2: **Per-head Δ_{primacy} (SFT − Base) across 32 heads × 36 layers.** Left: lowest-Σ run (IFD <1%); center: highest-Σ run (IFD 10%); right: mean across all six runs. All panels are cross-subset averaged (three disjoint 120-question subsets). Dotted vertical line marks the upper-half boundary (layer 18); shaded band marks the mid-layer reversal zone (layers 13–15). The color scale is symmetric around zero; left and center panels share a common scale, while the right panel uses a compressed scale for the six-run mean.

M Headline Statistic Provenance

Table M1 maps the main-text headline statistics to the specific runs, released artifacts, and statistical correction families from which they are derived. These artifacts are included in the released evaluation archive and anonymized repository.[3]

[3] https://anonymous.4open.science/r/3GPP_MCQ-7AB4

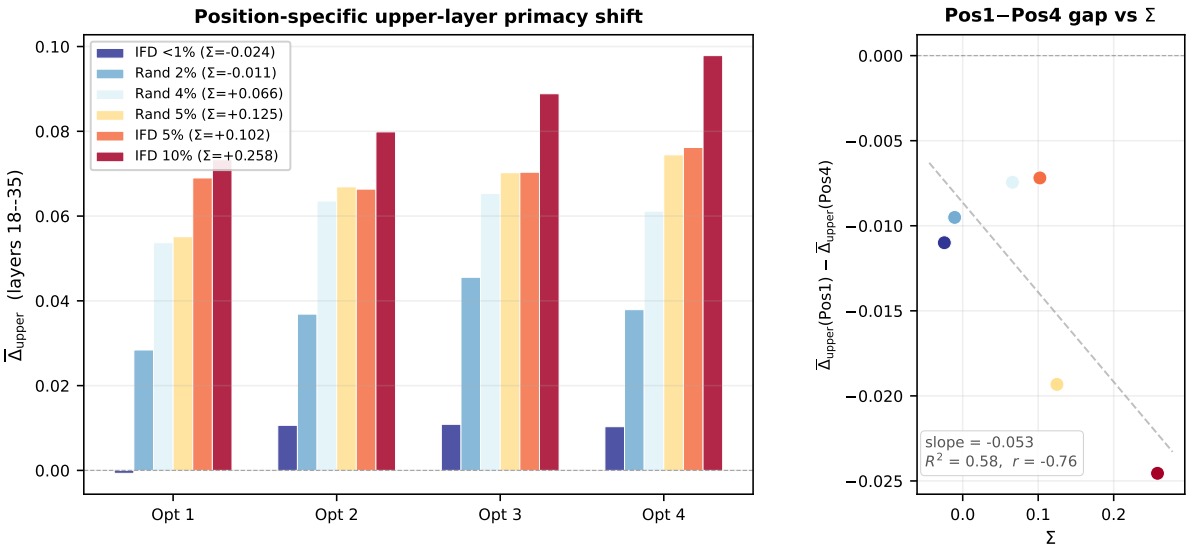

Figure K3: **Position-specific upper-layer primacy shift.** Left: $\overline{\Delta}_{\text{upper}}$ (layers 18–35) grouped by correct-answer position for each of six Qwen3-4B runs, pooled across three disjoint 120-question subsets (90 questions per position per run). Right: the Pos 1−Pos 4 gap in $\overline{\Delta}_{\text{upper}}$ plotted against Σ (slope $= -0.053$, $r = -0.76$); the gap widens with Opt 1 drift, consistent with the logprob-level accuracy model (Eq. 1).

Table L1: **Claim support mapping.** Each row connects a main-text evidence thread to the finding it supports, the primary evidence, the diagnostic or statistical test, and the boundary conditions under which the finding holds.

Evidence thread	Main finding	Primary evidence	Test / diagnostic	Boundary conditions
Separability	Position-averaged MCQ accuracy and predicted-position behavior can separate under matched evaluation	MedMCQA paired Open-QA/MCQ contrast: identical debiased accuracy (51.8% each, McNemar p=1.00) but a 14.8 pp MaxPred% gap (Fig. 1a; App. J.4)	McNemar exact; paired format contrast; predicted-position distribution	One medical-domain paired contrast; one seed per variant; supervision structure and token budget both differ; supports separation, not a causal mechanism
Target-domain scale	In the 3GPP primary campaign, SFT shifts predicted-position behavior in a structured L1a pattern, and a large share of raw score change is not stable to option-order averaging	36/40 runs exceed base MaxPred% (§5.1); $R^2 \geq 0.73$, LOO MAE 3.7 pp, 5-fold CV MAE 3.7 pp with $R^2_{\mathrm{OOS}} \geq 0.70$ (Table 2, Fig. 2, App. G); raw-to-debiased gap ratio 39% [33%, 46%] and all-run 41% [35%, 47%] (§5.3, Fig. 3)	OLS regression; repeated k-fold CV; bootstrap 95% CI; n=4→24 convergence check	Main quantitative claims use Qwen3-4B 40-run primary set; Σ is an L1a-specific directional summary; raw-to-debiased ratio is operational, not a causal fraction
Cross-architecture check	The L1a sign pattern transfers qualitatively, but slope magnitudes remain architecture-specific	Asymmetric signs replicate on Qwen, Mistral, and LLaMA; cross-prediction MAE 4.4–6.7 pp and cosine ≥ 0.915 (Table F1, Table F3, App. F)	Per-architecture OLS; sign consistency; slope-vector cosine; cross-prediction MAE	Mistral-Inst (9) and LLaMA-Inst (7) are smaller checks; Mistral differs in internal vulnerability ordering; supports qualitative transfer rather than a shared numerical model
Method-comparison consequence	Matched-budget, position-aware reporting weakens support for the tested IFD-over-Random claim on the homogeneous 3GPP pool	L1a debiased IFD–Random gap +1.64 pp (p_{Bonf}=0.182); L1b gap +0.2 pp (p_{Bonf}=1.000) (§6.1); control sequence in Fig. 4; Random 6-seed σ_{db}=1.89 pp, selector-side $\sigma \leq 0.53$ pp (App. A.5)	McNemar exact; Bonferroni k=12; Cohen's h; seed-variance analysis	Single-source 3GPP pool under matched controls; seed-level coverage for IFD/T-SHIRT; single-run matched coverage for PPL/RDS+/IFD+FLA; non-detection allows small advantages below the power floor
Pool-composition boundary	Selector separation re-emerges when inter-domain structure is introduced, so the homogeneous-pool result should not be generalized to all selector settings	Heterogeneous 50/50 pool: IFD–Random mean +5.5 pp (σ=3.3, 4 seeds, all Bonf-sig at k=4); bridge residual positive at all 8 selector–seed combinations; 25–75% ratio extension: 12/12 positive (binomial p<0.001), IFD $\sigma_{\mathrm{db}} \leq 0.6$ pp and Random σ_{db} 2.1–3.9 pp (Table J5, App. J.2)	McNemar exact; bootstrap 95% CI; Bonferroni k=4; binomial direction check; bridge decomposition	One 3GPP+Alpaca mixture family; L1a is the main endpoint; L1b/L2/MMLU gaps remain small; boundary evidence rather than a direct replacement for the homogeneous-pool comparison
Auxiliary mechanism probes	Format, flip-flow, and head-level analyses are consistent with position-mass redistribution but remain observational	Format ablation (App. H); correct-to-wrong flip flow (App. D); upper-layer attention to Opt 1 co-varies with Σ (Spearman ρ=0.94 on 3 disjoint subsets; Table K1, Fig. K1)	Descriptive ablation; flow analysis; Spearman rank	Qwen-focused; head probing covers n=6 checkpoints; no causal intervention on attention heads or FFN components

Table M1: **Provenance of headline statistics.** Run names match Table B1; released artifacts are included with the evaluation archive; correction families are in Appendix E.

Claim (§)	Run(s)	Eval	Released artifact	Corr.
MaxPred% excess: base +6.72→mean +17.61 pp, range +4.84–+43.20 pp (§5.1)	40 primary	L1a raw	`*/l1a_v32_logprob.json`	—
$R^2 \geq 0.73$, MAE 3.7 (§5.2)	40 primary	L1a raw	`generate_paper_tables.py` → `tab2_position_model.tex`	—
5-fold CV MAE 3.7 (App. G)	40 primary	L1a raw	`eq1_cross_validation.py` → `eq1_cross_validation.json`	—
39% raw-to-debiased gap ratio (§5.3)	33 w/ $\lvert\Delta\rvert \geq 1$ pp	raw+db	`bootstrap_ci_results.json` (filtered)	boot.
41% all-run (§5.3)	40 primary	raw+db	`bootstrap_ci_results.json` (all)	boot.
Random 80.00% db (§6.1)	Random 5% lr=2e−5, ep10	L1a db	`random5_lr2e5_ep10/l1a_v32_debiased.json`	—
IFD 81.64% db (§6.1)	IFD 5% lr=2e−5	L1a db	`ifd5_lr2e5/l1a_v32_debiased.json`	—
IFD 4% 83.52% db (§6.1)	IFD 4% lr=2e−5	L1a db	`ifd4_lr2e5/l1a_v32_debiased.json`	—
IFD–Rand p=.182 (§6.1)	IFD vs Rand ep10	L1a db	`*_debiased_details.json` → `compute_36test_matrix.py`	k=12
IFD–Rand p=.073 (§6.1)	IFD vs Rand ep10	L1a raw	`*_logprob_details.json` → `compute_36test_matrix.py`	k=12
Het-S mean +5.5 pp (§6.2)	4 seeds × 3 selectors	L1a db	`hetero_multiseed_results.json`; `sigma_bridge_4seed.json`	k=4
Ratio sweep 12/12 positive (§6.2)	4 seeds × 3 het. ratios + 1 hom.	L1a db	`tables/b1_ratio_replication.json`	binomial; k=4
Σ trajectory (Fig. J1)	2 IFD + 2 Rand, 75% 3GPP	L1a raw	`tables/b3_sigma_trajectory.json`	—
α_k cross-arch (§5.4)	55 Qwen + 9 MI + 7 LI	L1a raw	`cross_arch_position_analysis.py` → App. F	—
Slope transfer MAE (§5.4)	40 Qwen → MI + LI	L1a raw	`cross_arch_slope_transfer.py`	—
10/12 non-sig (§5.4)	3 arch × 4 eval	all	`cross_arch_mcnemar.json`	k=12
LLaMA disc 67/31 (§5.4)	LLaMA: IFD vs Rand 5%	L1a db	`*_details.json`	k=12
Qwen disc 41/21 (§5.4)	Qwen: IFD vs Rand 5% (ep10)	L1a db	`*_details.json`	k=12
Probing ρ=0.94 (§8)	6 matched Qwen	L1a subsets	`attention_analysis/head_probing_summary.json`	—