Alignment or Formatting?
Dissecting What On-Policy DPO Actually Learns

Anonymous ACL submission

Abstract

On-policy Direct Preference Optimization is commonly evaluated using held-out pairwise accuracy and LLM-judge win rate, but these metrics can reward different behaviors. In controlled Qwen3-4B ablations, interventions that improve a log-prob preference probe do not necessarily improve judged generations, and interventions that win generation judging need not improve the probe. The clearest gains in judge win rate come from direction correction, which strongly improves generation evaluation while leaving pairwise accuracy nearly unchanged. Output-level analysis shows that this advantage is driven primarily by response organization: the judged-winning model wins formatting on 40/50 prompts but wins instruction-following on only 11/50. Stripping markdown does not remove the advantage, suggesting that the effect is not merely markdown decoration but broader response organization. These results suggest that in this setting, on-policy DPO can optimize judge-preferred response organization without corresponding gains on a held-out preference probe.

1 Introduction

Preference optimization reports two numbers as if they were interchangeable: whether a model ranks held-out preference pairs correctly, and whether an LLM judge prefers its open-ended generations. Direct Preference Optimization (DPO; Rafailov et al., 2023) and its variants have made this practice common because both numbers look like preference alignment.

The two numbers can disagree (Figure 1). Under one Qwen3-4B (Qwen Team, 2025) recipe, self-play on-policy data improves the held-out log-prob probe by +0.008, while student-negative on-policy data (Chen et al., 2026) hurts it by -0.016 despite 95% prompt overlap between the two source pools. A filtering intervention then improves the probe by +0.022, but loses generation evaluation. A direction-correction intervention does the reverse: it changes the probe by only +0.003 across five seeds, but wins generation with length-controlled win rate 0.87 under GPT-5.4 and 0.95 under Kimi-K2 (Kimi Team, 2025).

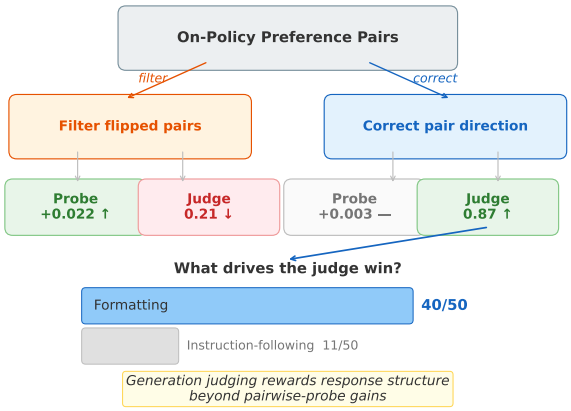

Figure 1: Overview. Filtering and direction correction produce opposite results on the held-out preference probe and LLM-judge generation evaluation.

The generation win is not a clean helpfulness win. A GPT-5.4 classification of 50 AlpacaEval 2 (Dubois et al., 2024) prompts shows that the direction-corrected model wins formatting on 40/50 prompts. The original-direction model wins instruction-following on 25/50 prompts and completeness on 23/50 prompts (remaining prompts are ties). The judge-preferred model is therefore the more structured model, not the model that better follows instructions or completes the task on this diagnostic.

This paper argues that on-policy DPO can improve judge-preferred response organization without corresponding gains on a held-out preference probe. The claim is bounded to the controlled setting studied here: Qwen3 models, LoRA (Hu et al., 2021) rank 64, UltraFeedback (Cui et al., 2023) preferences, student-negative and self-play on-policy data, and LLM-judge generation evalu-

ation. The point is methodological. A preference-data intervention can look successful under a judge because it teaches better information structure and response shape, even after markdown is removed.

The paper makes three contributions.

Controlled paired ablation. We show that filtering and direction correction move held-out preference probes and judged generation in opposite directions under matched controls.

Output-level diagnosis. A 50-prompt classification reveals that formatting is the dominant driver of the judge-preferred generation win: 40/50 prompts favor the direction-corrected model on formatting, while instruction-following and completeness favor the original-direction model.

Cross-loss and cross-scale checks. SimPO (Meng et al., 2024) and Qwen3-8B experiments show that the dissociation is not a one-loss or one-scale artifact.

2 Background

Preference optimization turns pairwise labels into a training signal for an assistant model. For prompt x, preferred response y^+, rejected response y^-, policy π_θ, and reference policy π_{ref}, DPO optimizes

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \tfrac{\pi_\theta(y^+|x)}{\pi_\theta(y^-|x)} - \log \tfrac{\pi_{\text{ref}}(y^+|x)}{\pi_{\text{ref}}(y^-|x)}\big]\Big). \tag{1}$$

The scalar β controls how strongly the loss reacts to the labeled preference margin. SimPO removes the reference model and optimizes a related margin objective (Meng et al., 2024). Both losses are used as direct alternatives to reward-model training and PPO-style RLHF (Stiennon et al., 2020; Ouyang et al., 2022; Schulman et al., 2017).

On-policy preference data is attractive because it is close to the model being updated. A model samples candidate answers, a judge labels the better answer, and the result becomes new preference data. Self-play data compares two model samples. Student-negative data keeps a stronger teacher answer and replaces the negative side with a student generation, following the broad motivation of distillation-style alignment (Chen et al., 2026). These pipelines are now common in iterative preference optimization (Xiong et al., 2024; Guo et al., 2024; Chen et al., 2024).

Two evaluation paradigms are usually mixed together. The first is a held-out pairwise probe: score each answer by conditional log probability and ask whether the preferred answer receives the higher score. The second is generation evaluation: sample responses to open-ended prompts and ask an LLM judge which model is better (Zheng et al., 2023; Wang et al., 2024). The probe tests ranking of fixed answers. The judge tests outputs the model chooses to write.

The implicit assumption is that both paradigms measure the same preference quality. That assumption is fragile. A model can rank fixed pairs better while generating less judge-preferred answers. A model can also win judge preference because it writes in a more structured format, even when a diagnostic judges another output better for instruction-following and completeness.

Related work. DPO and its variants are well studied (Azar et al., 2024; Ethayarajh et al., 2024; Tang et al., 2024; Chowdhury et al., 2024). On-policy and AI-labeled data are used widely (Bai et al., 2022; Cui et al., 2023; Tunstall et al., 2023; Argilla, 2024). LLM judges are known to be sensitive to position, length, and style (Zheng et al., 2023; Wang et al., 2024; Li et al., 2024), and DPO is sensitive to label noise and β choices (Xu et al., 2024; Wu et al., 2024; Ma et al., 2025; Wang et al., 2025). LoRA makes controlled multi-seed ablations feasible (Hu et al., 2021; Dettmers et al., 2023). This work connects these threads by showing that a preference probe and a generation judge reward different properties of the same on-policy training intervention.

3 Experimental Setup

The experiments isolate the training intervention while holding the model, budget, and evaluation protocol fixed. The main model is Qwen3-4B (Qwen Team, 2025) trained with LoRA rank 64 (Hu et al., 2021). The main loss is sigmoid DPO. Cross-loss checks replace it with SimPO. Cross-scale checks use Qwen3-8B.

Data. The base preference mixture comes from UltraFeedback Binarized (Hugging Face H4, 2023), derived from UltraFeedback (Cui et al., 2023). The self-play source samples two Qwen3-4B responses per prompt and asks a Kimi K2-family judge (Kimi Team, 2025) to choose between them. The student-negative source keeps a teacher-preferred answer, replaces the rejected side with a Qwen3-4B generation, and then asks the same judge to re-adjudicate the pair. A student-negative pair is flipped when

the judge prefers the response that the original label rejected.

Evaluations. The preference probe uses 1200 held-out pairs. Each answer is scored by answer-mean conditional log probability with prompt tokens masked. A pair is correct when the preferred answer receives the higher mean score. This is a held-out preference probe over fixed responses, not an open-ended helpfulness test.

Generation evaluation uses 805 AlpacaEval 2 prompts. Responses are judged by GPT-5.4 and Kimi-K2. Each pair is judged in both A/B orders, and the reported generation metric is length-controlled win rate. The length control matters because a judge can prefer longer or more structured responses for reasons that are not task accuracy.

Ablations. The design separates three possible explanations. First, matched-budget and same-prompt controls test whether self-play and student-negative data differ only because they cover different prompts. Second, a filtered student-negative control removes flipped pairs and compares it to a size-matched unfiltered control. Third, a paired direction-correction control keeps the same prompts and the same two responses, then swaps only the chosen/rejected labels. If preference direction were the main learned signal, the third contrast should move both the probe and generation. It does not.

4 Results

The controlled ablations show that the log-prob probe and LLM-judge generation preference reward different signals.

4.1 Two On-Policy Sources, Opposite Signs

Two prompt-overlapping on-policy sources move the preference probe in opposite directions. Table 1 gives the main matched-budget and same-prompt controls.

Self-play data improves the held-out probe by +0.008. Student-negative data hurts it by -0.016. The prompt pools overlap by 95%, so the sign difference is not a prompt-distribution artifact. The same-prompt control keeps the input set fixed and still preserves the gap. A beta sweep (not tabulated) gives the same message: low β is necessary, but calibration alone does not make the harmful student-negative source useful.

Table 2 shows that locality does not explain the sign. Both self-play locality splits remain positive.

Panel A: Shared-prompt control (318 prompts)			
Condition	**Seed 42**	**Seed 123**	**Mean**
Self-play high-loc	**0.6400**	**0.6567**	**0.6483**
Student-neg. high-loc	0.6225	0.6342	0.6283
Gap	**+0.0175**	**+0.0225**	**+0.0200**
Panel B: Matched-budget transfer (3 seeds)			
Condition	**Mean**	**SD**	**Δ vs base**
Base mix	0.6325	0.0038	+0.0000
Self-play high-loc	**0.6403**	0.0019	**+0.0078**
Student-neg. high-loc	0.6169	0.0102	**−0.0156**

Table 1: Matched-budget and shared-prompt controls. Panel A fixes the prompt set; Panel B uses three seeds with matched on-policy budgets.

Condition	**Mean**	**Seed SD**
Base mix	0.6325	0.0038
Self-play high-loc764	**0.6403**	0.0019
Self-play low-loc764	0.6389	0.0067
Student-neg. high-loc764	0.6169	0.0102
Student-neg. low-loc764	0.6100	0.0071
Student-neg. confirmed300	**0.6406**	0.0034
Student-neg. unconfirmed300	0.6186	0.0046
Self-play high-loc764 − base	**+0.0078**	0.0039
Self-play low-loc764 − base	+0.0064	0.0067
Student-neg. high-loc764 − base	**−0.0156**	0.0064
Student-neg. low-loc764 − base	**−0.0225**	0.0087
Student-neg. confirmed300 − base	**+0.0081**	0.0053
Student-neg. unconfirmed300 − base	**−0.0139**	0.0051

Table 2: Multi-seed validation (seeds 42, 123, 456). Both self-play splits are positive; both unfiltered student-negative splits are negative.

Both unfiltered student-negative locality splits remain negative. The failure is tied to what the pairs teach, not just where the prompts come from.

4.2 Filtering Helps the Probe, Not the Generator

Filtering flipped student-negative pairs improves the held-out preference probe but does not improve open-ended generation. Table 3 shows the probe-side contrast.

The filtered control beats the size-matched unfiltered control by +0.022 on the held-out probe, with a confidence interval that excludes zero. This result would look like a clean preference-data improvement if the probe were the only metric.

Generation gives the opposite answer. On AlpacaEval 2, the filtered model loses to the size-matched unfiltered model after length control. The length-controlled win rate for the filtered model is 0.21, with a three-seed bootstrap interval of [0.17,

Condition	Mean	Delta vs base
Base mix	0.6325	+0.0000
Student-neg. confirmed300	**0.6406**	**+0.0081**
Student-neg. unconfirmed300	0.6186	**−0.0139**
Student-neg. high-loc764	0.6169	**−0.0156**

Table 3: Size-matched filtering control (3 seeds). The filtered subset is positive while the unfiltered control remains negative.

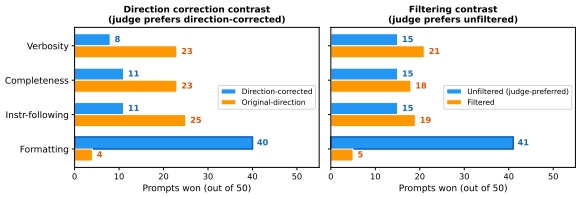

Figure 2: GPT-5.4 output classification on 50 AlpacaEval 2 prompts. Response structure dominates both contrasts; ties are omitted from counts.

0.25] — far below the 0.5 indifference point. Filtering therefore improves ranking of fixed held-out pairs while hurting judged generation.

4.3 Direction Helps Generation, Not the Probe

Direction correction wins generation while leaving the held-out probe almost unchanged. The paired direction control keeps the same 300 prompts and the same two responses per prompt, then swaps only the chosen/rejected labels. On the log-prob probe, the direction-corrected model beats the original-direction model by only +0.003 across five seeds, with 95% CI [-0.007, +0.013]. This rules out a probe effect comparable to the +0.022 filtering gain (per-seed data in Appendix Table 4).

The same checkpoints behave very differently as generators. Across three seeds, the direction-corrected model averages length-controlled win rate 0.87 under GPT-5.4 and 0.95 under Kimi-K2. The three-seed bootstrap interval under GPT-5.4 is [0.83, 0.90] (full table in Appendix Table 5). Direction correction is therefore the intervention that makes judges prefer the generated answers, even though it does not teach the model to rank held-out preference pairs much better.

4.4 Response Structure Drives the Generation Win

The judge-preferred generation win is driven by response organization rather than instruction quality.

Figure 2 shows a GPT-5.4 classification of 50 AlpacaEval 2 comparisons along four dimensions. The direction-corrected model wins formatting on 40/50 prompts. The original-direction model wins instruction-following on 25/50 and completeness on 23/50. The symmetric filtering contrast points the same way: the unfiltered model wins formatting on 41/50 prompts.

To test whether this is merely surface-level markdown, we strip all markdown symbols and re-judge (Appendix Table 9). The no-ties win rate barely changes (0.849 → 0.854 for direction; 0.206 → 0.247 for filtering). The judge-preferred quality is broader response organization that survives markdown removal.

4.5 Cross-Loss Validation: SimPO Shows the Same Dissociation

Changing the loss changes which intervention affects the probe, but not the metric dissociation. Under SimPO (Appendix Table 8), filtering gives -0.003 while direction correction gives +0.008 and is significant. The pattern is not identical to sigmoid DPO. That is the point. A data intervention is not a universal property of the pairs. It interacts with the loss, and the probe still does not explain the formatting-driven generation win.

4.6 Cross-Scale: 8B Models Show Direction Growing

Scaling to Qwen3-8B (Appendix Table 7) weakens filtering and strengthens direction correction on the probe. The filtering effect shrinks to +0.002. The direction effect grows to +0.007. This trend is consistent with the main story: the smaller DPO model benefits from removing bad pairs for the fixed probe, while the direction signal becomes more usable at larger scale. Generation evaluation at 8B is left as a limitation, so the scale result is a probe-side robustness check rather than a full reproduction of the formatting diagnosis.

4.7 Discussion

The probe is sensitive to pair composition; generation judging rewards broader response organization that survives markdown removal. These two metrics should not be treated as interchangeable.

5 Conclusion

In this setting, on-policy DPO can improve judge-preferred response organization without corresponding gains on a held-out preference probe. The advantage survives markdown removal. Held-out pairwise accuracy and LLM-judge win rate should not be treated as interchangeable.

Limitations

The strongest evidence comes from Qwen3-4B with LoRA rank 64, sigmoid DPO, and UltraFeedback-style preference data. Qwen3-8B provides a cross-scale log-prob check, but generation evaluation at 8B is not included. The SimPO result shows loss dependence rather than universal behavior.

The generation evidence uses LLM judges, not human raters. GPT-5.4 is independent of the Kimi-family training judge, while Kimi-K2 is not. The output classification also uses GPT-5.4, so it is a diagnostic for the judge preference rather than a replacement for human helpfulness evaluation.

Formatting is not inherently bad. Structured answers can be easier to read. The limitation is interpretive: a win caused by formatting should not be reported as a win in instruction-following, completeness, or general helpfulness without separate evidence.

The paired direction probe has five seeds, but per-pair held-out model outputs were not retained for a stronger paired bootstrap over the 1200 evaluation pairs. The main generation bootstrap covers the paired AlpacaEval setting, but it does not remove the need for human validation.

References

Argilla. 2024. DPO-Mix-7k: A mixed-source preference dataset. https://huggingface.co/datasets/argilla/dpo-mix-7k.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2024. A general theoretical paradigm to understand learning from human feedback. *International Conference on Artificial Intelligence and Statistics*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Xiwen Chen, Jingjing Wang, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Hejian Sang, Zhipeng Wang, Alborz Geramifard, and Feng Luo. 2026. SODA: Semi on-policy black-box distillation for large language models. *arXiv preprint arXiv:2604.03873*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. 2024. Provably robust DPO: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guo Tong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. In *International Conference on Machine Learning*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. *Advances in Neural Information Processing Systems*, 36.

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

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. *International Conference on Machine Learning*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Ber, and 1 others. 2024. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hugging Face H4. 2023. UltraFeedback Binarized. https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized. Dataset card.

Kimi Team. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment. *International Conference on Learning Representations*.

Qinwei Ma, Jingzhe Shi, Can Jin, Jenq-Neng Hwang, Serge Belongie, and Lei Li. 2025. Gradient imbalance in direct preference optimization. *ArXiv*, abs/2502.20847.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre H. Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *ArXiv*, abs/2402.05749.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yunan Wang, Jijie Li, Bo wen Zhang, Liangdong Wang, and Guang Liu. 2025. Inco-dpo: Balancing distribution shift and data quality for enhanced preference optimization. *ArXiv*, abs/2503.15880.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. β-dpo: Direct preference optimization with dynamic β. *ArXiv*, abs/2407.08639.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. *International Conference on Machine Learning*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? a comprehensive study. *International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Additional Experimental Details

A.1 Data, Judges, and Splits

Base data. The public source corpus is UltraFeedback Binarized (Hugging Face H4, 2023), derived from UltraFeedback (Cui et al., 2023). The base preference-mixture baseline uses 764 high-margin pairs from the processed training split and 764 Qwen3-8B-judged pairs. The 1200-pair evaluation set is from the processed test split.

Self-play source. For each prompt, Qwen3-4B (Qwen Team, 2025) samples two responses (temperature 0.8, top-p 0.95). `Kimi-K2-0905-jcloud` (Kimi Team, 2025) adjudicates the pair. The pool contains 2296 pairs from 1394 prompts.

Student-negative source. Following Chen et al. (2026), the student-negative pool keeps the teacher chosen response and replaces the rejected with a Qwen3-4B generation. The 1419 pairs are re-adjudicated by the same Kimi judge. 54.5% are flipped.

Split rules. The locality proxy is $0.7 \cdot \mathrm{sim} + 0.3 \cdot (1 - \mathrm{gap})$. Self-play splits use confidence ≥ 0.5; student-negative splits use ≥ 0.8. The filtered 300-pair control removes flipped rows. The unfiltered control keeps the same budget without removing flips.

A.2 Paired Direction-Correction Probe Data

Table 4 gives per-seed held-out accuracies for the paired direction-correction ablation described in Section 4.3. Both rows use the same 300 prompts and the same two responses per prompt; only the chosen/rejected labels differ. The three-seed paired Δ is +0.0031 with 95% CI [-0.022, +0.028]. Adding seeds 789 and 2024 (not tabulated) narrows the five-seed CI to [-0.007, +0.013], ruling out direction effects comparable to the +0.022 filtering gain.

A.3 AlpacaEval 2 Generation Tables

Table 5 reports the full AlpacaEval 2 dual-judge generation results referenced in Section 4.2 and Section 4.3. Each prompt is judged in both A/B orders by GPT-5.4 (independent of training judge)

Condition	Seed 42	Seed 123	Seed 456	Mean
FlipKimi300	0.6125	0.6017	0.6117	0.6086
FlipOrig300	0.6050	0.6100	0.6017	0.6056

Table 4: Per-seed held-out accuracies for the paired direction-correction ablation. Same 300 prompts and response pairs in both rows.

and Kimi-K2 (same family as the training adjudicator). Agreed-direction wins are counted. Length-controlled win rates (LC-WR) come from the logistic fit $\Pr(A \succ B) = \sigma(\beta_0 + \beta_1 \ell)$, where $\ell = (\text{len}_A - \text{len}_B)/(\text{len}_A + \text{len}_B)$. The joint three-seed bootstrap CIs are reported in the caption.

A.4 Generated-Output Classification

To diagnose *why* the judge prefers the direction-corrected model, we classify 50 randomly sampled AlpacaEval 2 prompts (seed 42 checkpoint) on four dimensions using GPT-5.4: verbosity, completeness, formatting, and instruction-following. Table 6 shows that formatting dominates both contrasts. In the direction contrast, the direction-corrected model wins formatting on 40/50 prompts but loses instruction-following on 25/50 and completeness on 23/50. In the filtering contrast, the unfiltered model wins formatting on 41/50. This confirms that LLM-judge preference tracks response structure rather than task quality in these comparisons.

A.5 Cross-Scale Replication

Table 7 replicates the core contrasts on Qwen3-8B with three seeds. The filtering effect shrinks from +0.022 on Qwen3-4B to +0.002 on Qwen3-8B, while the direction effect grows from +0.003 to +0.007. Neither Qwen3-8B contrast excludes zero at $\text{df} = 2$, so this should be read as a scaling trend rather than a definitive replication. Generation evaluation at 8B is left to future work.

A.6 Cross-Loss Replication: SimPO

Table 8 replaces sigmoid DPO with SimPO (Meng et al., 2024) (β=2.0, γ=1.0, reference-free) on the same Qwen3-4B data and splits. Under SimPO, filtering becomes slightly negative (-0.003) while direction correction becomes the larger and statistically significant effect (+0.008). This cross-loss reversal confirms that which intervention dominates the log-prob probe is loss-dependent, not a universal property of the preference pairs.

A.7 Format-Normalization Experiment

To test whether the generation advantage is driven by literal markdown tokens, we strip all markdown symbols (headings, bullets, bold, italic, code blocks, horizontal rules) from both sides of each comparison and re-judge with GPT-5.4 under the same bidirectional protocol.

Judge	A vs B	Raw WR	LC-WR	β_{len}	p	len_A/len_B
GPT-5.4	Conf300 vs Unconf300	0.206	0.216	+0.25	0.642	0.64
Kimi-K2	Conf300 vs Unconf300	0.045	0.069	+2.67	3.0e-4	0.64
GPT-5.4	FlipKimi vs FlipOrig (s42)	0.849	0.857	-0.25	0.628	1.65
GPT-5.4	FlipKimi vs FlipOrig (s123)	0.860	0.852	+0.20	0.705	1.82
GPT-5.4	FlipKimi vs FlipOrig (s456)	0.833	0.889	-1.25	0.028	2.06
Kimi-K2	FlipKimi vs FlipOrig (s42)	0.969	0.947	+3.02	1.3e-5	1.65
Kimi-K2	FlipKimi vs FlipOrig (s123)	0.978	0.961	+2.48	0.004	1.82
Kimi-K2	FlipKimi vs FlipOrig (s456)	0.969	0.930	+3.15	2.3e-5	2.06

Table 5: AlpacaEval 2 dual-judge results. Joint 3-seed bootstrap CI for FlipKimi-vs-FlipOrig: [0.83, 0.90] (GPT-5.4), [0.92, 0.97] (Kimi-K2). Joint CI for Conf-vs-Unconf: [0.17, 0.25] (GPT-5.4), [0.06, 0.12] (Kimi-K2).

	FlipKimi vs FlipOrig		Conf vs Unconf	
Dimension	A wins	B wins	A wins	B wins
Verbosity	8	23	21	15
Completeness	11	23	18	15
Formatting	**40**	4	5	**41**
Instruction-following	11	25	19	15

Table 6: GPT-5.4 output classification ($n = 50$, ties omitted). Formatting dominates both contrasts. Direction contrast: A = direction-corrected, B = original-direction. Filtering contrast: A = filtered, B = unfiltered.

Model	H+AI	Conf−Unconf	FlipK−FlipO	Seeds
Qwen3-4B	0.6325	**+0.0220**	+0.0030	3
Qwen3-8B	0.5789	+0.0020	**+0.0072**	3

Table 7: Cross-scale log-prob replication. Filtering shrinks and direction grows at 8B. Neither Qwen3-8B contrast excludes zero at $\mathrm{df} = 2$.

Loss	H+AI	Conf−Unconf	FlipK−FlipO	Seeds
Sigmoid DPO	0.6325	**+0.0220**	+0.0030	3
SimPO	0.5658	−0.0033	**+0.0075**	3

Table 8: Cross-loss replication on Qwen3-4B. Under SimPO, filtering becomes negative while direction correction becomes the larger effect. Bold marks the dominant intervention per loss.

Contrast	Original WR (no-ties)	Stripped WR (no-ties)
Direction (FlipKimi vs FlipOrig)	0.849	0.854
Filtering (Conf vs Unconf)	0.206	0.247

Table 9: Format-normalization results (GPT-5.4, seed 42). Stripping markdown barely changes the no-ties win rate in either contrast, indicating that the judge preference is not driven by markdown tokens but by broader response organization.