

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

Keyu Wen Yang Song Lijun Yang Rui Lin

JD.com, China

{wenkeyu1, songyang.0461, cdyanglijun, linrui}@jd.com

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 often reports two numbers as if they measured the same property: whether a model ranks fixed held-out preference pairs correctly, and whether an LLM judge prefers its opened generations. DPO (Rafailov et al., 2023) and related objectives make both measurements easy to obtain, but they exercise different model behaviors.

Figure 1 previews the dissociation. Under one Qwen3-4B (Qwen Team, 2025) recipe, self-play on-policy data improves the held-out probe by +0.008, whereas prompt-overlapping student-negative data (Chen et al., 2026) reduces it by -0.016. More decisively, filtering re-adjudicated pairs whose preferred direction flips improves the probe by +0.022 but loses generation evaluation. A paired *direction-correction* control holds its 300

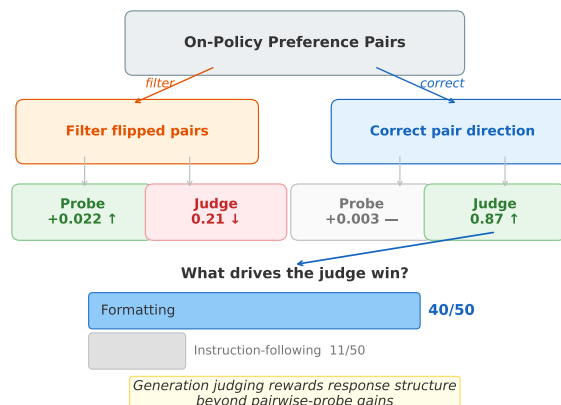


Figure 1: Overview. Filtering and direction correction produce opposite results on the held-out preference probe and LLM-judge generation evaluation.

prompts and both candidate responses fixed, changing only which response is marked preferred; it moves the probe by only +0.003 across five seeds but wins generation.

The judge win is not automatically a helpfulness win. A GPT-5.4 diagnosis of 50 AlpacaEval 2 (Dubois et al., 2024) comparisons assigns the direction-corrected model a formatting win on 40/50 prompts, whereas the original-direction model wins instruction-following on 25/50 and completeness on 23/50. A blind human diagnostic below does not simply agree with this LLM diagnosis. We therefore treat every score as a separate measurement, not as ground truth for general helpfulness.

Our claim is methodological and bounded: Qwen3 models, rank-64 LoRA (Hu et al., 2021), UltraFeedback-derived preferences (Cui et al., 2024), student-negative/self-play data, and LLM-judge evaluation. It is not a recipe for improving alignment. Instead, it shows why a preference-data intervention that appears successful under one metric should not be interpreted as a general quality gain without another evaluation view.

The paper makes three contributions.

Controlled paired ablation. Filtering and direction correction move the held-out probe 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

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y^+|x)}{\pi_\theta(y^-|x)} - \log \frac{\pi_{\text{ref}}(y^+|x)}{\pi_{\text{ref}}(y^-|x)} \right] \right). \quad (1)$$

The scalar β controls how strongly the loss reacts to a labeled margin. SimPO removes the reference model (Meng et al., 2024); both objectives are alternatives to reward-model training and PPO-style RLHF (Stiennon et al., 2020; Ouyang et al., 2022; Schulman et al., 2017).

On-policy data is close to the model being updated: a model samples answers, then a judge labels a pair. Self-play compares two model samples. *Student-negative* data retains a teacher-preferred answer and replaces the rejected side with a student generation (Chen et al., 2026). Such iterative, AI-labeled pipelines are common (Xiong et al., 2024; Guo et al., 2024; Chen et al., 2024).

Our held-out *preference probe* scores fixed answers by conditional log-probability and asks whether the preferred response receives the higher normalized score. Generation evaluation instead samples new answers and asks an LLM judge to choose between them (Zheng et al., 2023; Wang et al., 2024). The former tests ranking; the latter tests what the model writes.

Related work. DPO variants, theoretical treatments, and robust objectives study how a loss uses preference labels (Gheshlaghi Azar et al., 2024;

Ethayarajh et al., 2024; Tang et al., 2024; Chowdhury et al., 2024). Work on LLM-as-a-judge identifies position, verbosity/length, and style sensitivity (Zheng et al., 2023; Wang et al., 2024; Li et al., 2024); DPO studies identify label-noise and β sensitivity (Xu et al., 2024; Wu et al., 2024; Ma et al., 2025; Wang et al., 2025). We connect these literatures by testing whether the same intervention transfers across a fixed-pair probe, LLM judging, and a human diagnostic.

3 Experimental Setup

The experiments vary one data intervention while fixing model, budget, and evaluation. The main model is Qwen3-4B (Qwen Team, 2025), trained with sigmoid DPO and LoRA (Hu et al., 2021); SimPO and Qwen3-8B are boundary checks.

Data. The base mixture is UltraFeedback Binarized (Hugging Face H4, 2023), an AI-feedback-derived dataset, not a human-preference ground truth. Self-play samples two Qwen3-4B responses and asks a Kimi K2-family judge (Kimi Team, 2025) to choose. Student-negative data replaces a teacher pair’s rejected response with a Qwen3-4B response, then re-adjudicates it. A pair is *flipped* when this judge reverses the original direction. *Filtering* keeps the non-flipped pairs; *direction correction* instead holds prompts and responses fixed while using the re-adjudicated direction.

Evaluations. The preference probe uses 1,200 held-out pairs. Prompt tokens are masked and answers are ranked by answer-mean conditional log probability. Generation uses 805 AlpacaEval 2 prompts, GPT-5.4 and Kimi-K2, both A/B orders, and ties for order-inconsistent verdicts. We report a length-controlled win rate (LC-WR) for Qwen3-4B; it separates a judge’s choice from response-length differences.

Ablations. Unless stated otherwise, training uses rank 64, $\alpha=128$, dropout 0.05, learning rate 2×10^{-5} , two epochs, batch size 4, gradient accumulation 4, maximum length 512, and $\beta=0.002$. Main controls use seeds 42/123/456; the direction probe additionally uses 789/2024. The design separates source, filtering, and direction. Matched-budget and same-prompt controls test whether self-play and student-negative data differ only because they cover different prompts. The 300-pair filtering control removes flipped student-negative rows and

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.

compares them with a size-matched unfiltered subset. The paired direction control instead keeps the same prompts and response texts, changing only the chosen/rejected assignment. The appendix gives split rules, decoding, judge prompts, and all result rows.

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. 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

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.

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.

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. Its length-controlled win rate is 0.216 under GPT-5.4 and 0.069 under Kimi-K2, with three-seed bootstrap intervals [0.17, 0.25] and [0.06, 0.12], respectively. 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

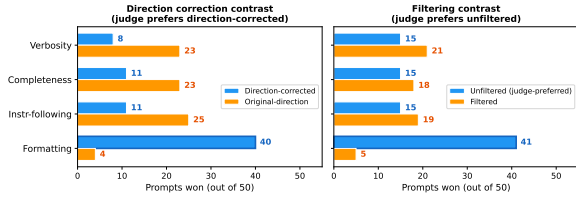


Figure 2: GPT-5.4 output classification on 50 AlpacaEval 2 prompts. Response structure dominates both contrasts; ties are omitted from counts.

gain (per-seed data in Appendix Table 7).

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 8). 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 mark-down, we strip all markdown symbols and re-judge (Appendix Table 13). The no-ties win rate barely changes (0.849 $\rightarrow$ 0.854 for direction; 0.206 $\rightarrow$ 0.247 for filtering). The judge-preferred quality is broader response organization that survives mark-down removal.

Prompt and human robustness. Under content-first and format-neutral prompts, the direction-corrected model still wins: GPT-5.4 gives no-tie rates 0.794/0.800 and Kimi-K2 gives 0.960/0.968; filtering remains disfavored under both judges (Table 4). A blinded one-rater diagnostic on 100 comparisons finds 43/45/12 wins/losses/ties for direction-corrected versus original (no-tie WR 0.489), rather than reproducing the LLM-judge win. This is diagnostic evidence of metric dissociation,

A vs B	Judge	Content-first	Format-neutral
Direction-corrected vs original	GPT-5.4	0.794	0.800
Direction-corrected vs original	Kimi-K2	0.960	0.968
Filtered-300 vs unfiltered-300	GPT-5.4	0.275	0.253
Filtered-300 vs unfiltered-300	Kimi-K2	0.068	0.051

Table 4: No-tie, order-balanced win rates under content-oriented judge rubrics ($n = 805$ prompts). The qualitative direction is unchanged.

A vs B (dimension)	A	B	Tie	A WR
Corrected vs original (overall)	43	45	12	0.489
Corrected vs original (instruction)	33	18	49	0.647
Corrected vs original (completeness)	43	21	36	0.672
Corrected vs original (organization)	39	50	11	0.438
Filtered vs unfiltered (overall)	46	40	14	0.535
Filtered vs unfiltered (instruction)	18	27	55	0.400
Filtered vs unfiltered (completeness)	23	44	33	0.343
Filtered vs unfiltered (organization)	50	35	15	0.588

Table 5: Blind one-rater human diagnostic (100 comparisons per contrast). WR excludes ties and maps left/right answers back to hidden conditions.

not a human-preference benchmark.

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 12), 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. A matched-size source control likewise gives 0.6115 with no extra data, 0.6248 with on-policy data, and 0.6340 with off-policy data (Appendix Table 6); on-policy origin alone does not determine the outcome.

4.6 Cross-Scale: 8B Models Show Direction Growing

Scaling to Qwen3-8B (Appendix Table 11) weakens filtering and strengthens direction correction on the probe. The filtering effect shrinks to +0.002. The direction effect grows to +0.007. We additionally evaluate generation on 805 AlpacaEval 2 prompts for three seeds: GPT-5.4 no-tie win rates are 0.886/0.853/0.868 (mean 0.869) for direction correction and 0.148/0.217/0.191 (mean 0.185) for filtering. These 8B rates are order-balanced but not length-controlled, so they complete a qualitative cross-scale generation check rather than a full dual-judge replication.

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 controlled Qwen3 DPO, preference-direction correction can substantially improve LLM-judge generation preference without a commensurate fixed-pair probe gain, while filtering can do the opposite. The result survives two content-oriented judge rubrics and recurs in an 8B generation check, but the blind human diagnostic does not simply mirror it. We therefore recommend reporting held-out preference probes, generation judging, and human evaluation as distinct measurements, together with their judges, prompts, tie rules, and length treatment. The evidence is diagnostic rather than a general claim about all on-policy or preference-optimization methods.

Limitations

The strongest evidence comes from Qwen3-4B with rank-64 LoRA, sigmoid DPO, and UltraFeedback-style AI feedback. We add a Qwen3-8B generation check, but only GPT-5.4 order-balanced no-tie rates are available there; it is not a full length-controlled, dual-judge replication. SimPO changes the probe pattern, demonstrating loss dependence rather than universal behavior.

The generation evidence uses LLM judges. GPT-5.4 is independent of the Kimi-family training judge, whereas Kimi-K2 is not. The output classification also uses GPT-5.4. Our blind human diagnostic has one rater and no inter-rater-agreement estimate; it demonstrates disagreement among metrics, not population-level human preference.

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 1,200 evaluation pairs. The main generation bootstrap covers the paired AlpacaEval setting.

References

- Xiwen Chen, Jingjing Wang, Wenhui Zhu, Peijie Qiu, Xuanzhao Dong, Yueyue Deng, 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, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [ULTRA FEEDBACK: Boosting language models with scaled AI feedback](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR.
- 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*.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. [A general theoretical paradigm to understand learning from human preferences](#). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. 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, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 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.
- 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. [\$\beta\$ -dpo: Direct preference optimization with dynamic \$\beta\$](#) . *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, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Additional Experimental Details

A.1 Data, Judges, and Splits

Training recipe. Unless noted in the main text, sigmoid DPO uses Qwen3-4B, LoRA rank 64, $\alpha = 128$, dropout 0.05, learning rate 2×10^{-5} , two epochs, per-device batch size 4, gradient accumulation 4, maximum length 512, and $\beta = 0.002$. Main runs use seeds 42, 123, and 456. The paired direction probe additionally uses seeds 789 and 2024. SimPO uses $\beta = 2.0$ and $\gamma = 1.0$ without a reference model.

Base data. The public source corpus is UltraFeedback Binarized (Hugging Face H4, 2023), derived from UltraFeedback (Cui et al., 2024). 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 $0.7 \cdot \text{sim} + 0.3 \cdot (1 - \text{gap})$. 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.

Generation-judge prompt. The standard generation evaluation uses our custom pairwise prompt, not an AlpacaEval leaderboard/default prompt: “Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user instruction below. Choose the assistant that follows the instruction better, considering helpfulness, relevance, accuracy, depth, creativity, and overall quality. Avoid position bias. Do not let response length dominate the judgment. Output exactly one of: [[A]], [[B]], or [[C]] (tie) as your final verdict.” It then presents the instruction and the two delimited responses. For the robustness

Setting	On-policy	Off-policy
Matched size, $\beta = 0.002$	0.6248	0.6340
Calibrated rank-64	0.6515	0.6527

Table 6: Source and calibration boundary controls. The no-extra-data baseline in the matched-size setting is 0.6115.

prompts, *content-first* asks the judge to prioritize correctness, instruction following, coverage, and factual usefulness while ignoring style, headings, bullets, markdown, and organization unless they change substance. *Format-neutral* asks it to mentally remove markdown, headings, bullet layout, numbering, and code fences before judging helpfulness, completeness, accuracy, and usefulness. All three prompts require exactly [[A]], [[B]], or [[C]]; both A/B orders are evaluated at temperature zero.

A.2 Prompt Sensitivity and Human-Evaluation Protocol

For prompt sensitivity, the original AlpacaEval 2 response pairs are re-judged in both A/B orders. *Content-first* emphasizes correctness, instruction following, coverage, and factual usefulness; *format-neutral* asks the judge to ignore superficial organization unless it changes substance. Order-inconsistent outcomes are ties. Table 4 in Section 4.4 reports no-tie win rates for A, where A is the first condition named.

One blinded annotator independently selected left, right, or tie for 100 AlpacaEval responses in each contrast; condition labels were hidden. The results in Table 5 in Section 4.4 are a diagnostic only: a single annotator cannot estimate inter-rater reliability or population preference.

A.3 Source and Calibration Boundary

Table 6 tests whether the main result is explained simply by adding on-policy data. In the matched-size $\beta = 0.002$ control, off-policy data slightly exceeds on-policy data. With the calibrated rank-64 setting ($\beta = 0.002$, $\beta_{\text{op}} = 0.0005$), the two sources are nearly tied. These controls delimit the claim: source quality and calibration, rather than the on-policy label alone, determine the probe outcome.

A.4 Paired Direction-Correction Probe Data

Table 7 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.

Condition	Seed 42	Seed 123	Seed 456	Mean
Direction-corrected	0.6125	0.6017	0.6117	0.6086
Original-direction	0.6050	0.6100	0.6017	0.6056

Table 7: Per-seed held-out accuracies for the paired direction-correction ablation. Same 300 prompts and response pairs in both rows.

A.5 AlpacaEval 2 Generation Tables

Table 8 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) 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.6 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 9 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.7 Cross-Scale Replication

Table 11 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-side

judging is available below; its values are order-balanced no-tie rates rather than LC-WR.

A.8 Cross-Loss Replication: SimPO

Table 12 repeats the probe with SimPO (Meng et al., 2024): filtering falls to -0.0033 , whereas direction correction rises to +0.0075, showing that the probe result is loss-dependent.

A.9 Format-Normalization Experiment

Table 13 reports a markdown-stripping control, re-judged by 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	Filtered-300 vs unfiltered-300	0.206	0.216	+0.25	0.642	0.64
Kimi-K2	Filtered-300 vs unfiltered-300	0.045	0.069	+2.67	3.0e-4	0.64
GPT-5.4	Direction-corrected vs original (s42)	0.849	0.857	-0.25	0.628	1.65
GPT-5.4	Direction-corrected vs original (s123)	0.860	0.852	+0.20	0.705	1.82
GPT-5.4	Direction-corrected vs original (s456)	0.833	0.889	-1.25	0.028	2.06
Kimi-K2	Direction-corrected vs original (s42)	0.969	0.947	+3.02	1.3e-5	1.65
Kimi-K2	Direction-corrected vs original (s123)	0.978	0.961	+2.48	0.004	1.82
Kimi-K2	Direction-corrected vs original (s456)	0.969	0.930	+3.15	2.3e-5	2.06

Table 8: AlpacaEval 2 dual-judge results. Joint 3-seed bootstrap CI for direction-corrected-vs-original: [0.83, 0.90] (GPT-5.4), [0.92, 0.97] (Kimi-K2). Joint CI for filtered-vs-unfiltered: [0.17, 0.25] (GPT-5.4), [0.06, 0.12] (Kimi-K2).

Dimension	Corrected vs original		Filtered vs unfiltered	
	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 9: 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.

A vs B	s42	s123	s456	Mean
Corrected vs original	0.886	0.853	0.868	0.869
Filtered-300 vs unfiltered-300	0.148	0.217	0.191	0.185

Table 10: Qwen3-8B GPT-5.4 generation check on 805 prompts per seed. Entries are order-balanced no-tie win rates for A, without length control.

Model	H+AI	Filtered—unfiltered	Corrected—original	Seeds
Qwen3-4B	0.6325	+0.0220	+0.0030	3
Qwen3-8B	0.5789	+0.0020	+0.0072	3

Table 11: Cross-scale log-prob replication. Filtering shrinks and direction grows at 8B. Neither Qwen3-8B contrast excludes zero at $df = 2$.

Loss	H+AI	Filtered—unfiltered	Corrected—original	Seeds
Sigmoid DPO	0.6325	+0.0220	+0.0030	3
SimPO	0.5658	−0.0033	+0.0075	3

Table 12: 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 (corrected vs original)	0.849	0.854
Filtering (filtered vs unfiltered)	0.206	0.247

Table 13: 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.