ECHO: EMERGENT COMMON-MODE HIDDEN OFFSETS DRIVE VOCABULARY COLLAPSE IN PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization can improve every ranking metric while building a hidden, shared representation shift that the relative loss barely sees, one that can eventually destroy generation. We trace this failure to the *common-mode offset*, a shared directional shift in the model's hidden states that appears across token positions, prompts, and both chosen and rejected responses. Because it moves chosen and rejected representations together, the relative ranking loss barely reacts to its accumulation. As this offset grows, it projects onto the output layer to form a *vocabulary well*, elevating a single token's probability until it dominates decoding and causes generation collapse. We demonstrate the causal role of this mechanism through targeted interventions. The isolated offset alone reconstructs the vocabulary well, injecting the learned offset into an untreated base model recreates the collapse where norm-matched random directions do not, and masking this common component from the loss prevents the well from forming. Under the per-token ranking loss we study, the same collapse appears in 14 of 17 base-model lineages and under full-parameter fine-tuning. Moreover, the offset is not merely an artifact of collapse, since it is already present in textbook DPO and SimPO runs that form no well. Textbook DPO keeps it weakly aligned with any single vocabulary direction even with three times our base budget. Trained from base models for that budget, however, full-parameter SimPO drives it into collapse on both Llama-2 and Llama-3.1, and the same kind of offset grows from a public SFT checkpoint under the official SimPO recipe. Across 17 UltraFeedback runs, the better a run ranks held-out responses, the less often two public reward models prefer its samples to the base model's. These findings expose a blind spot in standard ranking metrics and introduce a representation-level signal that can warn of generation collapse before it occurs.

1 INTRODUCTION

Preference optimization has become the standard paradigm for adapting pretrained language models toward human-preferred responses (Ouyang et al., 2022), ranging from pairwise Direct Preference Optimization (DPO) (Rafailov et al., 2023) to listwise objectives based on Plackett–Luce ranking (Plackett, 1975; Luce, 2005; Liu et al., 2025). Recent pipelines push this ranking signal further, with repeated rounds of preference optimization on continually refreshed preference data (Yuan et al., 2024; Grattafiori et al., 2024). However, recent work reveals a growing disconnect between successful ranking optimization and healthy text generation (Rafailov et al., 2024). Razin et al. (2025) identify *likelihood displacement*, where the absolute likelihood of the chosen response can plummet even as its relative preference margin improves. Ren & Sutherland (2025) show that suppressing rejected tokens can undesirably squeeze probability mass toward tokens the base model already favored. Yuan et al. (2025) mathematically explain why chosen and rejected likelihoods move together via *gradient entanglement*. While these studies illuminate how probabilities shift during ranking, the terminal state of this dynamic remains largely unexplained. How does this movement manifest inside the model, why does it eventually fixate on a seemingly irrelevant token, and why do standard ranking metrics look healthy while generation breaks down?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

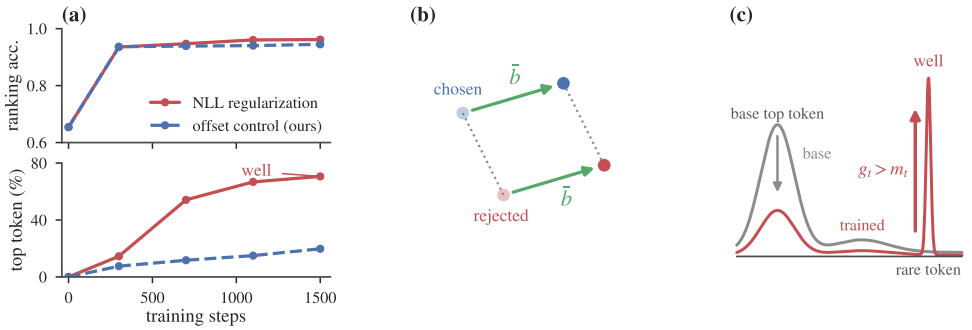

Figure 1: (a) Two Llama-2 runs on the same driving lists rank held-out responses equally well, but only the one without offset control collapses onto one token. (b) The common-mode offset $\bar{b}$ moves chosen and rejected hidden states together, so the margin the loss sees barely changes. (c) The offset lifts a rare token above the base model's top token once its gain g_t exceeds the base gap m_t.

We find that preference training induces a surprisingly simple internal anomaly. The final hidden states begin to drift in nearly the identical direction across token positions, prompts, and both chosen and rejected responses. We term this shared movement the *common-mode offset*. Because this offset acts as a constant added to both chosen and rejected responses, a ranking loss that optimizes their relative difference has almost no incentive to penalize it. Consequently, this shared shift can keep growing in the background, masked by improving preference margins.

This hidden shift becomes destructive because the output layer projects these hidden states into vocabulary logits. When the offset pushes a single token's logit far higher than the tokens the base model would predict, it creates a *vocabulary well*. Once this well becomes deep enough to dictate decoding, the model enters *generation collapse*, uncontrollably repeating that single token. Importantly, the well need not be a rare token with an unusually large output-vector norm, since the same mechanism produces a well on the ordinary token `Conf` on UltraFeedback (Cui et al., 2024).

Causal interventions confirm that this offset is the mechanism of collapse. Keeping only its common component reconstructs the well, injecting it into an untreated base model recreates the collapse where norm-matched random directions do not, and centering the rejected-side loss to hide it prevents the well (§5). The offset moves both chosen and rejected scores by tens of nats but their margin by less than 3 nat in an average list, which is why held-out ranking improves while generation collapses (Figure 1). Across 17 UltraFeedback runs, the better a run ranks held-out responses, the less often two public reward models prefer its samples to the base model's (§8).

This vulnerability is not an artifact of one unusual hyperparameter setting. Our main objective is the ranking term of PRO (Song et al., 2024), and we observe the same collapse signature in 14 of 17 base-model lineages and under full-parameter fine-tuning (§4). Moreover, the common-mode offset grows under every objective we test, including textbook DPO and SimPO, and the objectives differ in how hard they push it. We measure this push by the loss's *demand*, the per-token margin beyond which it stops pushing. Textbook DPO at its usual setting has a low demand, so its offset stays weakly aligned with any vocabulary direction and generation stays intact. This protection comes from the low demand rather than from the reference model, since adding a reference model to the per-token objective does not prevent the well (§6). SimPO lies between the two. Trained from base models for three times our base budget, full-parameter SimPO drives the offset into collapse, and the offset also grows from a public SFT checkpoint under the official SimPO recipe.

Therefore, generation collapse is not an inevitability of preference optimization, but the extreme terminal state of a hidden dynamic that is active long before generation visibly fails. Penalizing the offset's norm and its projection onto the vocabulary prevents collapse up to three times the base budget, and at longer budgets the same decomposition shows where the remaining pressure goes.

Prior work showed that preference optimization can move probability mass where it should not go and couples chosen and rejected updates. We trace these probabilistic symptoms to one translation vector in the final hidden state that the loss barely sees, show that it causes the well, and predict its onset. This exposes a blind spot in ranking metrics and gives a measurable health signal.

2 Related Work

Likelihood displacement, squeezing, and gradient entanglement. Razin et al. (2025) demonstrate that DPO training can decrease the chosen response's absolute likelihood and predict *when* a dataset risks it from the geometry of the unembedding matrix. Yuan et al. (2025) prove that margin-based losses entangle the gradients of chosen and rejected log-probabilities, explaining *why* the two tend to move in tandem. While these works map the symptoms in probability space, we isolate their source in representation space, the *common-mode offset* in the final hidden state. Furthermore, Ren & Sutherland (2025) show that the negative gradient on rejected tokens "squeezes" probability mass toward tokens the base model is already confident about. The vocabulary wells we discover are different, since the base model initially ranks our well tokens far below its preferred tokens, typically by 17 to 39 nat, until the offset inflates their logits to dominate decoding.

Regularizing absolute likelihood. DPO-Positive (Pal et al., 2024), an explicit NLL term (Pang et al., 2024), and length-normalized regularization toward the reference (Im et al., 2026) constrain scalar probabilities rather than the shared direction in the hidden state. In our experiments, an NLL term on the chosen response delays the well without preventing it (§7).

Unembedding geometry and representation anisotropy. Pretrained embeddings and hidden states share a common mean, a few dominant directions, and outlier dimensions (Mu & Viswanath, 2018; Ethayarajh, 2019; Biś et al., 2021; Gao et al., 2019; Puccetti et al., 2022; Macocco et al., 2025; Sun et al., 2024; Stolfo et al., 2024), whereas the common-mode offset is built during preference *training*. Functionally, it resembles activation steering (Rimsky et al., 2024), and DPO is known to change behavior by shifting the residual stream (Lee et al., 2024), but here preference optimization grows the steering vector on its own until generation collapses. Many of our severe wells emerge from under-trained tokens (Land & Bartolo, 2024), which merely provide a path of least resistance for the offset, and decoding-time corrections (Ding et al., 2026) leave the training-time cause in place. Appendix A gives more detail.

3 Setup

Listwise and pairwise preference objectives. Our main runs train the model to rank a list of responses. Given a prompt x and K responses ranked from most to least preferred, $y^{(1)} \succ y^{(2)} \succ \cdots \succ y^{(K)}$, the sequential Plackett–Luce objective (Plackett, 1975; Luce, 2005) asks the model to prefer the top remaining response over all responses below it, one stage at a time, by minimizing

$$\mathcal{L}_{\text{PL}} = -\sum_{k=1}^{K-1} \log \frac{\exp(\beta\, s_\theta(x, y^{(k)}))}{\sum_{j=k}^{K} \exp(\beta\, s_\theta(x, y^{(j)}))}, \qquad s_\theta(x, y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}).$$

Each response is scored by the average per-token log-probability of the policy π_θ, without a reference model, and $\beta > 0$ sets the scale of the score differences. Up to this scale, the objective recovers the ranking term of PRO (Song et al., 2024), which also adds a supervised term on the top response. For $K = 2$ it becomes the length-normalized, reference-free pairwise objective of SimPO (Meng et al., 2024) without its target margin γ, which we call *pairwise ranking*.

We train on two preference datasets. The first comprises 5,383 driving-preference lists adapted from Li et al. (2025), featuring four responses ranked by the severity of their physical consequences. The second is pairwise UltraFeedback data (Cui et al., 2024). As a boundary case we also train textbook DPO (Rafailov et al., 2023), which scores a whole response by its summed log-probability ratio against a frozen base model, $s_\theta(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$. We primarily train LoRA adapters (Hu et al., 2022), optionally adding an NLL term $\lambda_{\text{NLL}} \cdot \text{NLL}(y^{(1)})$ on the chosen response (Pang et al., 2024), and some runs on Llama-2 and Llama-3.1 use full-parameter fine-tuning.

A nearly unconstrained translation. Every objective above depends only on differences between scores. Adding the same amount to the scores of all responses of a prompt therefore leaves the loss unchanged. This invariance makes rewards under the Bradley–Terry (Bradley & Terry, 1952) and Plackett–Luce models identifiable only up to a prompt-dependent constant (Rafailov et al., 2023). A shift of the hidden states shared by every position does not add exactly the same amount to every score, since its effect on each token depends on the token and its context. To first order, however, the

objective sees such a shift only through its component along one direction, a weighted difference of the responses' mean score gradients (Appendix B.17). Whether a learned shift lies mostly outside that direction is an empirical question, which §5 answers by measurement.

The common-mode offset. Let h_t and h_t^{base} denote the final hidden states that the unembedding reads at position t of a rejected response (evaluated under teacher forcing) for the trained and base models, respectively. We define the displacement as $\Delta h_t = h_t - h_t^{\text{base}}$. The *common-mode offset* $\bar{b}$ is the mean displacement over the response's positions, $\bar{b} = \frac{1}{T}\sum_{t=1}^{T} \Delta h_t$, averaged again over a probe set of rejected responses. The *residual* $r_t = \Delta h_t - \bar{b}$ represents the position-specific part, such that $\Delta h_t = \bar{b} + r_t$ holds exactly. This decomposition requires only two checkpoints and no access to training internals. We track the norm $\|\bar{b}\|$, the cosine between the offsets of chosen and rejected responses, and the *loudness* $\kappa \log \sum_v \exp\big((W\bar{b})_v/\kappa\big)$, a smooth maximum of its projection onto the vocabulary, where W is the unembedding matrix and κ a temperature.

Three stages of failure. *Likelihood displacement* is the routine erosion of absolute log-likelihoods for both chosen and rejected responses and, in isolation, is not pathological. *Vocabulary well formation* is the intermediate state, in which the probability mass released by the displacement concentrates on one or a few anomalous tokens or token families; it can be measured before any text is sampled. *Generation collapse* is the terminal state, in which these tokens dominate decoding. We count a run as *collapsed* once it reaches either state, that is, once one token takes most of the released mass or most of its samples degenerate, by repeating text or by cycling through the forms of one word. Released mass can also spread thinly across the vocabulary without forming a well. We call the resulting loss of sample quality *diffuse degradation*, measure it by the base model's per-token negative log-likelihood of the samples, and do not count it as collapse.

Measurement and configuration. *Mass flow* is computed under teacher forcing on rejected responses. At every position we track the probability mass surrendered relative to the base model, the *released mass*, the share of it absorbed by the top-1 token, the *capture*, and the share leaking beyond the 10,000th-ranked vocabulary item, the *tail*; a well has *nucleated* once it captures the majority of the released mass. *Degeneration* scores sampled generations for n-gram repetition (Holtzman et al., 2020; Welleck et al., 2020), together with a word-stem ring rate and the base model's per-token negative log-likelihood of the samples, which detects non-repeating degeneration that standard repetition counts overlook (Appendix B.14). *Held-out ranking* is evaluated on 600 unseen lists, or on 300 of them for a few runs, by top-1 accuracy, pairwise accuracy, and the *margin* between the chosen response's mean log-probability and that of the highest-ranked rejected response (Appendix B.2). Unless stated otherwise, we fine-tune a rank-32 LoRA adapter with a learning rate of 10^{-5} and an effective batch size of 32; full-parameter runs use a learning rate of 10^{-6}. A *base budget* is three passes over the training data, which is 504 steps on the driving lists and about three quarters of one epoch of the official SimPO training data on UltraFeedback. Extended runs operate at fixed multiples of this budget under the same learning-rate schedule, and a base model's *onset budget* is the smallest multiple at which its untreated run collapses. Offsets and mass flow use 80 rejected training responses unless stated otherwise.

4 Ranking Improves While Generation Collapses

We first train Llama-2-7B (Touvron et al., 2023b) with a LoRA adapter under sequential $K = 4$ Plackett–Luce ranking on the driving lists. The lists were built for the Llama-2-based driving agent of Li et al. (2025). On 600 unseen lists, top-1 accuracy reaches 0.995, and pairwise accuracy stays above 0.94. Generation, however, fails. Within 504 steps, 89% of the probability mass released relative to the base model concentrates on a single token, `textt`, which the base model ranked deep in its vocabulary tail, and almost every sampled generation degenerates into repeating it. By every ranking metric the objective has succeeded, yet the output distribution has collapsed.

We replicate this training across 17 base-model lineages from five tokenizer families, spanning 0.5B to 14B parameters. Fourteen of the 17 collapse within six times the base budget (Table 1; Appendix Table 4), including TinyLlama-1.1B and the Qwen2.5 models from 0.5B to 3B. Twelve form a single-token well, and in Qwen2.5-14B and Gemma-2-9B most samples cycle through the forms of one word, `collision` or `crash`. Phi-3-mini and Gemma-2-2B, tested only up to three times the

Table 1: Every base here but Llama-3.2-3B collapses, onto one token or the forms of one word. Untreated runs at their onset budget. Ring counts samples dominated by one word stem, on the 12 saved samples where marked † and unmeasured where marked –.

Base	Budget	Δchosen (nat)	Capture / tail	Degen. / ring	$\|\bar{b}\|$	Well
Llama-2-7B	1×	−10.97	89.4% / 89.5%	96.0% / –	92.4	`textt`
Mistral-7B-v0.3	1×	−17.51	95.1% / 95.1%	100% / 0%	300.4	`/******/`
Qwen2.5-7B	3×	−21.67	93.9% / 86.7%	99.3% / –	328.7	CJK fragment
Qwen2.5-14B	3×	−3.85	27.0% / 7.7%	80.3% / 78.7%	198.6	`collision`-fam.
Gemma-2-9B	1×	+0.62	12.5% / 37.5%	0% / 11 of 12†	163.7	`crash`-fam.
Llama-3.1-8B	1×	−14.27	87.6% / 88.5%	18.3% / 22.0%	115.2	`.scalablytyped`
OLMo-2-7B	6×	−2.90	89.6% / 89.6%	21.0% / 8.0%	87.8	`contador`
OLMo-3-7B	3×	−6.83	93.8% / 93.8%	96.0% / 0%	79.7	(quote+4×newline)
Llama-3.2-3B	6×	+1.17	43.7% / 8.3%	10.7% / 7.7%	66.7	none (ordinary tokens)

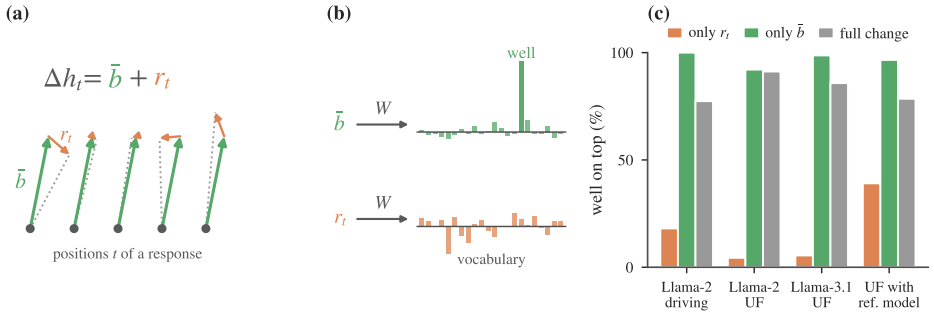

Figure 2: (a) At every position of a rejected response, the shift of the final hidden state splits into the shared offset $\bar{b}$ and a position-specific residual r_t. (b) Through the unembedding W, only $\bar{b}$ singles out one vocabulary row. (c) Adding only $\bar{b}$ to the base hidden state puts the well on top at nearly every rejected position, while adding only r_t rarely does; UF denotes UltraFeedback.

budget, stay below this criterion, although 39% of Phi-3-mini's samples repeat `speed`. The clear exception is Llama-3.2-3B. Its rejected responses fall by over 30 nat per token, yet its released mass stays on ordinary tokens and its samples stay nearly as fluent as the base model's (§6).

The collapse does not depend on adapters either. Full-parameter fine-tuning of Llama-2-7B at a learning rate of 10^{-6} collapses onto `textt` by step 250 of the same 504-step budget. Under full-parameter fine-tuning, onset tracks the learning rate summed over training, and from two earlier runs we wrote down before this run that it would collapse between steps 200 and 300 (Appendix B.10). The failure is therefore tied neither to one model lineage nor to adapters.

5 The Common-Mode Offset Causes the Well

The offset is shared across positions, prompts, and both responses. We computed the offset separately for each of 900 rejected responses and compared each with their global mean. Across prompts, the per-response offsets align with their mean with median cosines above 0.97 on each of the six runs we measured. Across the two sides of a preference pair, the chosen and rejected offsets exhibit cosines of 0.83 to 0.996. The shift is therefore a single geometric translation that moves both sides of every pair together (Figure 1b). Furthermore, this shared vector points at the well, since for single-token wells the well token has the largest entry of the projected vector $W\bar{b}$ (Figure 2b).

The loss barely sees the offset. We added the trained offset to the base model's hidden states and recomputed the scores of chosen and rejected responses on held-out lists. In the four collapsing runs we evaluated, the offset lowers both scores by 15 to 65 nat per token but moves the margin between them by only 0.2 to 2.7 nat in an average list (Appendix Table 8). Its effect on both scores is 20 to 530 times its effect on their difference, and its cosine with the gradient of that difference in the base model is at most 0.16. The objective therefore observes a healthy margin, while the log-probabilities underneath fall by tens of nats. A z-loss (Chowdhery et al., 2023; Team OLMo et al., 2024) that

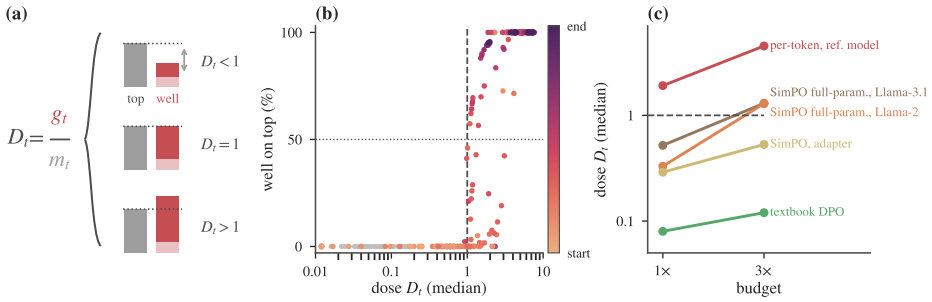

Figure 3: (a) The offset overrides the base model at a position once its gain g_t on the would-be well exceeds the base gap m_t, that is, once the dose D_t exceeds one. (b) Across 510 snapshots of ten collapsing lineages, colored by training progress, and three runs that stay healthy, in gray, the well never tops a majority of rejected positions while the median dose stays below one. (c) Median doses of Table 2 at the usual and the three-fold budget.

compresses the log-partition function 2.5-fold still leaves 65 to 66% of the released mass in the well (Appendix B.8). Nothing in the loss therefore holds the offset back, and fine-tuning alone already builds one. A control trained only with the NLL term grows an offset of norm 56 that chosen and rejected responses share (cosine 0.99), yet it lifts no token by more than 3.4 (Appendix B.11).

Keeping, hiding, adding, or removing the offset controls the well. We test the offset's causal role in four ways. The first adds $\bar{b}$ alone or r_t alone to the base hidden state. On untreated runs, $\bar{b}$ alone makes the well the top token at nearly every rejected position, more often than the full trained state, while r_t alone rarely does (Figure 2c). The residual therefore partly works against the well. The same holds under full-parameter fine-tuning, where the fine-tuned unembedding rows alone almost never lift the well (Appendix B.10).

The second test hides the offset from the loss. *Centering* computes the rejected-side ranking loss on $h_t - \text{stopgrad}(\bar{b})$. The rejected scores then no longer change when the offset grows, so pointing $\bar{b}$ at one vocabulary row no longer lowers them. The gradient can still move the offset, since it treats the subtracted mean as a constant. On the Llama-2 flagship, centering keeps most of the offset's norm but turns its projection away from the vocabulary, and no well forms.

The third test transplants the trained offset into an untreated base model. This *injection* adds $\bar{b}$ to the final hidden state at every position, causing the well token to dominate generation on every held-out prompt. Stripping only the well's row component from $\bar{b}$ removes that well, but generation instead collapses onto a different single token, whereas random or orthogonal directions of the same norm produce no well. The well therefore follows the offset rather than any particular token. The fourth test removes the offset from a model that has already collapsed. Subtracting half of its logit bias $W\bar{b}$ from the untreated flagship model at decoding time lowers the share of degenerate greedy samples on 60 held-out prompts from 100% to 0%, and the chosen response regains 8.7 nat per token (Appendix B.7). Ultimately, keeping only the offset reconstructs the well, hiding it prevents the well, adding it creates one, and removing it stops the collapse, establishing the offset as a controllable, sufficient cause of collapse.

6 When the Offset Becomes a Well

A well forms only after the offset's gain passes the base gap. The formation of a well depends on a logit ratio. Let m_t be the *base gap*, which measures how much the base model prefers its own top token over the candidate well token at position t, and let g_t be the logit advantage that $\bar{b}$ adds to that well token. In the offset-only state, the well overtakes the base model's top token where the *dose* $D_t = g_t/m_t$ exceeds one (Figure 3a). Scaling the injected offset confirms this threshold, since a quarter-strength $\bar{b}$ fails to trigger the well while a half-strength one pushes D_t above one and puts the well on top at nearly all rejected positions. We also tracked the dose along natural training runs (Figure 3b; Appendix Figure 9 shows each lineage). Across all 510 snapshots, the well never tops a majority of rejected positions while the median dose stays below one, so in our data a well needs a

dose above one, but a dose above one does not guarantee a well. In Llama-3.2-3B the median dose reaches 2.75 on the ordinary subword `e`. Yet the trained model puts `e` on top at only 29% of rejected positions, against 80 and 83% for the wells of Llama-3.1 and Llama-2, and its samples stay nearly as fluent as the base model's.

Wells tend to land on anomalous vocabulary rows. Because the same $\bar{b}$ projects more strongly onto rows with large norms, rare tokens with large unembedding rows gain the most. In the adapter runs on the driving data, most single-token wells sit in the top 1% of their base model's row norms and never occur in WikiText-103 (Merity et al., 2017), whereas in Llama-3.2-3B the highest-norm rows belong to common subwords. High row norms are not required, however. On UltraFeedback the Llama family collapses onto the ordinary token `Conf`, which the offset alone still puts on top at nearly every position (Appendix B.15). Row geometry makes some tokens more exposed, but the offset is what lifts them.

Different losses push the offset by very different amounts. We trained Llama-2-7B on the UltraFeedback pairs with our per-token ranking loss, the same loss with a reference model, SimPO, and textbook DPO, each at its usual setting and at three times the budget. We also ran SimPO with full-parameter fine-tuning on Llama-2-7B and Llama-3.1-8B. Table 2 and Appendix Figure 11 report each run's demand, offset norm, dose, capture, share of repetitive or nearly empty samples, and win rate over the base model under Skywork-Reward-V2 (Liu et al., 2026).

Only the per-token ranking loss collapses at the usual budget, and adding a reference model does not stop it. SimPO forms no well at its usual setting, although its samples still degrade, and it passes the threshold with three times the budget under full-parameter fine-tuning on both Llama-2 and Llama-3.1 (Figure 3c). Textbook DPO stays far below the threshold even with three times the budget, because its offset stays short and spreads across the vocabulary.

This difference stems from how far each loss keeps pushing. A ranking loss stops pushing once the chosen response beats the rejected one by a large enough margin, and we call the per-token size of that margin the loss's *demand*.[1] Our per-token loss keeps pushing until the gap reaches about 42 nat per token, well beyond the base model's own preference for its top token, so the dose passes one. SimPO stops at about 2.4 nat. Textbook DPO sums the score over the whole response, which spreads the same margin across hundreds of tokens and stops it at about 0.3 nat per token.

The reference model is therefore not what protects textbook DPO. Adding one to the per-token loss leaves the well in place on both datasets and at both budgets, with an offset as long as the untreated run's. What protects textbook DPO from the well is its low demand. Summing the score is equivalent to averaging it with β multiplied by the response length, so the protection comes from the resulting demand at the usual β rather than from the aggregation itself.

Between these extremes, SimPO within its published range grows an offset even longer than the collapsing run's, but a quiet one whose dose remains below one, so no well forms. Nevertheless, its generation quality degrades. The released mass spreads deep into the vocabulary instead of settling on one token, and the samples devolve into strings of unrelated words that standard repetition counts fail to detect, while the reward model prefers the base model's sample on about nine in ten prompts. However, adding the SimPO offset alone to the base model does not reproduce this diffuse form (Appendix B.13). This diffuse degradation is therefore a second route to damage, separate from the well. The missing fine-tuning stage alone does not explain this damage, since textbook DPO from the same base model wins against it on about two thirds of prompts.

Under a three-fold budget, however, the offset grows louder, concentrating its projection onto a narrow set of vocabulary rows. Its dose keeps rising, and under full-parameter fine-tuning it passes one on both Llama-2 and Llama-3.1. By the end of training, 80 to 86% of samples degenerate, and held-out ranking is the best of any UltraFeedback run on that base.

[1] A logistic ranking loss never stops pushing exactly, but once the scaled score margin reaches 5, its gradient weight has fallen to about 1% of its value at zero margin. We take this point as the stopping margin, which gives a demand of about $5/\beta$ nat under per-token averaging and $5/(\beta|y|)$ nat under scores summed over a response of $|y|$ tokens, with $|y| = 182$ on average in our UltraFeedback pairs. SimPO's target margin γ adds γ/β to its demand. Appendix Figure 10 shows the full sweep over β.

Table 2: Each loss at its usual setting and with three times the budget, with doses above one in bold. Runs start from the base model unless marked SFT, and all train on the same UltraFeedback pairs. SimPO uses $\beta = 2.5$ and $\gamma = 1.0$ from base models and the official $\beta = 2.0$ and $\gamma = 1.0$ from the SFT checkpoint. Degen. counts repetitive or nearly empty samples, and Win compares a run's samples with those of the model it starts from.

Objective	Budget	Demand (nat)	$\|\bar{b}\|$	Dose	Capture	Degen.	Win
Per-token ranking, $\beta = 0.12$	1×	42	84	**1.92**	97%	99%	25%
with a reference model	1×	42	79	**1.91**	86%	99%	26%
with a reference model	3×	42	91	**4.53**†	89%	99%	17%
SimPO	1×	2.4	114	0.29	0.4%	0.7%	7%
SimPO	3×	2.4	82	0.53	9%	28%	6%
SimPO, full-parameter	1×	2.4	65	0.33	3%	0.7%	10%
SimPO, full-parameter	3×	2.4	85	**1.29**	27%	80%	7%
on Llama-3.1-8B	1×	2.4	75	0.52	0.3%	30%	13%
on Llama-3.1-8B	3×	2.4	101	**1.30**†	45%	86%	3%
from SFT, on Llama-3-8B	1×	3.0	33	0.15	2.5%	1.3%	63%
from SFT, on Llama-3-8B	3×	3.0	77	0.19	2.7%	0%	34%
Textbook DPO, $\beta = 0.1$	1×	0.3	14	0.08	4%	3%	68%
Textbook DPO, $\beta = 0.1$	3×	0.3	44	0.12	4%	1%	66%

†Computed for the run's own leading token, `D` or `Apache`; all other doses are for `Conf`.

We also trained SimPO from its authors' public SFT checkpoint of Llama-3-8B with their hyperparameters (Appendix B.12). From this start, no well forms within three times the base budget, and the median dose of every token stays below 0.4. The offset nonetheless grows from a norm of 33 at the base budget to 77 at three times the budget. It stays quiet, lifting no token by more than 4.3 nat, against about 18 nat for SimPO from the base models at the same budget. Meanwhile the SFT model's per-token negative log-likelihood of the samples rises from 1.35 to 3.86, close to the 3.1 to 4.3 of the base-start SimPO runs (Appendix B.13). Two public reward models agree, preferring the one-fold run's samples to the SFT model's on 63 to 65% of held-out prompts but the three-fold run's on only 33 to 34%. Supervised fine-tuning enlarges the base gaps of the candidate wells by at most 22%, so by the dose it should delay the well rather than prevent it.

Supervised fine-tuning already produces an offset of the same form, and the preference loss sets where it points and how far it grows (Appendix B.11). Small β and listwise objectives therefore do not invent a new failure mode; they push harder on a shift that fine-tuning already produces.

7 Controlling the Offset

Centering hides the offset from the loss but leaves it intact within the model. Our recipe therefore keeps centering and adds two penalty terms that act on the offset itself, used alone or together, and we call the combination *both*. A norm penalty $\lambda_{\text{norm}}\|\bar{b}\|$ shrinks the offset, and a loudness penalty $\lambda_{\text{loud}} \cdot \kappa \log \sum_v \exp\big((W\bar{b})_v/\kappa\big)$ flattens its vocabulary projection without necessarily shrinking its norm. On the driving lists the recipe also keeps the NLL term of weight 1.0, so it differs from the NLL-only baseline only by offset control. Offset control matters most for objectives worth keeping despite a high demand. Listwise ranking, for example, also supervises the order among the rejected responses, which pairwise objectives never see, and on held-out driving lists it keeps that order at 0.82 to 0.92 while pairwise ranking inverts it (Appendix B.6).

Cross-family validation up to three times the base budget. We applied the recipe to four families at or beyond the budget where each untreated run collapses, with Llama-2, Qwen2.5-7B, and Llama-3.1 at three times the base budget (Table 3). No run forms a well. The largest share of the released mass, at most 20%, goes to an ordinary token such as the period, the tail stays at or below 5.5%, and sampled degeneration remains at or below about 12%. By contrast, at the same three-fold budget on Llama-2, the NLL-only baseline, N-DPO, and DPO-Positive leave 71 to 81% of the released mass in a single well (Table 3; Figure 1a; Appendix Figure 13). The cost to ranking performance is small, since held-out pairwise accuracy settles between 0.94 and 0.96. On Llama-2, Mistral, and Qwen2.5-7B, this is at most 1.5 points below a run of the same base and budget without offset control. Controlling the offset is therefore enough to prevent collapse up to three times

Table 3: The recipe, centering plus a norm penalty, a loudness penalty, or both, at or beyond each base's onset budget, where capture goes to an ordinary token. Three Llama-2 baselines without offset control keep 71 to 81% of the released mass on one rare token.

Base	Penalty	Steps	Capture / tail	Degen.	Held-out top-1 / pairwise
Llama-2-7B	loudness	1512	19.8% / 1.5%	6–12%	1.00 / 0.95
Mistral-7B	both	504	14.6% / 1.4%	3.0%	1.00 / 0.94
Qwen2.5-7B	both	1512	8.0% / 0.2%	1.7%	1.00 / 0.96
Llama-3.1-8B	both	1512	13.0% / 5.5%	6.7%	1.00 / 0.96
Llama-2-7B without offset control					
NLL only	–	1512	70.6% / 71.0%	2.0%	1.00 / 0.97
N-DPO	–	1512	78.1% / 78.4%	18.0%	1.00 / 0.96
DPO-Positive	–	1512	81.4% / 82.2%	24.3%	1.00 / 0.96

the base budget, at almost no cost to ranking. At a six-fold budget on the OLMo family, the residual rather than $\bar{b}$ comes to lift the surviving well, and penalizing it closes the remaining routes on OLMo-3 (Appendix B.4).

8 What Ranking Metrics Miss

An external judge disagrees with the ranking metrics. Held-out pairwise accuracy levels off at 0.94 to 0.96 within 300 to 500 steps, treated or not, while the margin keeps growing. The untreated run with the largest held-out margin typically collapses, in line with the over-optimization of direct alignment algorithms (Rafailov et al., 2024) (Appendix Figure 13). We evaluated one sample for each of 300 held-out UltraFeedback prompts using two public reward models, Skywork-Reward-V2 (Liu et al., 2026) and ArmoRM (Wang et al., 2024), across 17 Llama-2 runs (Appendix Table 7 and Figure 12a). Across these runs, the better a run's held-out pairwise accuracy, the less often the external judge prefers its sample to the base model's, with a Spearman correlation of -0.78 for Skywork-Reward-V2 and -0.79 for ArmoRM. The runs that rank best move the most probability away from the base model's likely tokens, and the judge tracks how much stays there. The share of mass remaining on the base model's ten most likely tokens correlates with the win rate at $+0.74$ and $+0.87$. Since none of our base models are instruction-tuned, we use the win rate only to compare runs. By this measure, the recipe with an NLL term of weight 1.0 matches textbook DPO, winning 65% of prompts against 66 to 68%.

The offset warns before the well forms. In seven of the nine lineages with snapshots every 50 steps, the eventual well enters the eight largest entries of $W\bar{b}$ before it takes over, up to 1,200 steps ahead, even while held-out ranking has long saturated (Appendix Figure 12b). Tracking $\|\bar{b}\|$, its loudness, and the leading entries of $W\bar{b}$ during training therefore offers an early signal that standard ranking metrics do not give, although we have read it only from completed runs.

9 Conclusion

Preference optimization can collapse generation while held-out ranking stays high. We traced this failure to a common-mode offset, a shift of the final hidden state shared across positions and across chosen and rejected responses, which the ranking objective barely penalizes. A well forms only after the offset lifts one token above the tokens the base model would otherwise predict. Different ranking losses turn the same preference pressure into very different amounts of the shift. Hiding the offset from the loss and penalizing it prevents collapse up to three times the base budget. Across runs, held-out ranking accuracy even anticorrelates with an external judge ($\rho = -0.78$). Ranking accuracy and preference margin measure only what the loss constrains, so the offset deserves monitoring.

Limitations. Our evidence spans 17 base-model lineages, two datasets, and adapter and full-parameter training, with two further seeds for the flagship settings and one seed elsewhere (Appendix B.16). The penalty strengths are calibrated on the driving data, the six-fold result covers the OLMo family only, and apart from the SFT-start runs the demand analysis starts from base models. We have not tested IPO (Gheshlaghi Azar et al., 2024), ORPO (Hong et al., 2024), or KTO (Ethayarajh et al., 2024).

Reproducibility Statement

Training configurations (base models, LoRA rank, learning rate, batch composition, and every penalty coefficient reported), the exposure-ordering and offset-decomposition measurement code, and the held-out ranking probe are described in §3–§7 and Appendix B; anonymized code and the training and held-out preference data will be released upon publication. Held-out ranking numbers are computed on 600 lists, or on 300 of them where stated, whose prompts and chosen responses do not appear in any training split; the corpus counts in §6 use the full training text and the WikiText-103 training split.

AI Use Statement

The authors set the research questions, made every design decision, and checked every reported number against the logged experimental runs and every citation against its source. Within this process, we used generative AI tools (Claude Code and Gemini) to help develop the conceptual framework, propose and refine hypotheses, design experiments and give feedback on the methodology, formulate mathematical claims and assist in writing their proofs, implement the methods, clean and reformat datasets, interpret results, and assist with translation. We did not use these tools to generate synthetic datasets. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv [cs.CL]*, 2024. URL https://arxiv.org/abs/2404.14219.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. Too much in common: Shifting of embeddings in transformer language models and its implications. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5117–5130, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.403. URL http://dx.doi.org/10.18653/v1/2021.naacl-main.403.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 0006-3444,1464-3510. doi: 10.2307/2334029. URL http://dx.doi.org/10.2307/2334029.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 2023. URL https://arxiv.org/abs/2204.02311.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9722–9744, 2024. doi: 10.48550/arXiv.2310.01377. URL http://dx.doi.org/10.5555/3692070.3692454.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. Breaking the likelihood trap: Variance-Calibrated Modulation for large language model decoding. *arXiv [cs.CL]*, 2026. doi: 10.48550/arXiv.2606.22511. URL http://arxiv.org/abs/2606.22511. Accepted at EMNLP 2026.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/d19-1006. URL http://dx.doi.org/10.18653/v1/D19-1006.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 2 February 2024. doi: 10.48550/arXiv.2402.01306. URL http://dx.doi.org/10.5555/3692070.3692574.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation Degeneration Problem in Training Natural Language Generation Models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://dblp.org/rec/conf/iclr/GaoHTQWL19.html.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv [cs.CL]*, 2024. doi: 10.48550/arXiv.2408.00118. URL http://arxiv.org/abs/2408.00118.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv [cs.AI]*, 2024. doi: 10.48550/arXiv.2407.21783. URL http://arxiv.org/abs/2407.21783.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.626. URL http://dx.doi.org/10.18653/v1/2024.emnlp-main.626.

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

Shawn Im, Federico Danieli, Skyler Seto, Barry-John Theobald, and Katherine Metcalf. Normalized rewards for preference optimization. In *Proceedings of the 43rd International Conference on Machine Learning*, Seoul, South Korea, 2026. doi: 10.48550/arXiv.2607.16240. URL http://dx.doi.org/10.48550/arXiv.2607.16240.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv [cs.CL]*, 2023. doi: 10.48550/arXiv.2310.06825. URL http://arxiv.org/abs/2310.06825.

M G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. ISSN 0006-3444,1464-3510. doi: 10.2307/2332226. URL http://dx.doi.org/10.2307/2332226.

Sander Land and Max Bartolo. Fishing for magikarp: Automatically detecting under-trained tokens in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11631–11646, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.649. URL http://dx.doi.org/10.18653/v1/2024.emnlp-main.649.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. In *International Conference on Machine Learning*, pp. 26361–26378. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24a.html.

Yun Li, Ehsan Javanmardi, Simon Thompson, Kai Katsumata, Alex Orsholits, and Manabu Tsukada. Multi-PrefDrive: Optimizing Large Language Models for Autonomous Driving Through Multi-Preference Tuning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4347–4354, 2025. URL https://ieeexplore.ieee.org/abstract/document/11247608/.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J Liu, and Xuanhui Wang. LiPO: Listwise preference optimization through learning-to-rank. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2404–2420, Stroudsburg, PA, USA, 2025. Association for Computational Linguistics. ISBN 9798891761896. doi: 10.18653/v1/2025.naacl-long.121. URL http://dx.doi.org/10.18653/v1/2025.naacl-long.121.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy. In *International Conference on Learning Representations*, volume 2026, pp. 133805–133838, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d8836d0294651a867c2ca51012923997-Abstract-Conference.html.

R D Luce. *Individual choice behavior: A theoretical analysis*. Dover Publications, New York, NY, 2005. ISBN 9780486441368. URL https://books.google.com/books?hl=en&lr=&id=D74qAwAAQBAJ&oi=fnd&pg=PP1&dq=Individual+Choice+Behavior:+A+Theoretical+Analysis&ots=1h7LEH7B-g&sig=x5QbofqnxfQbtctkLNX13ZbPjBM.

Iuri Macocco, Nora Graichen, Gemma Boleda, and Marco Baroni. Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. In Yonatan Belinkov, Aaron Mueller, Najoung Kim, Hosein Mohebbi, Hanjie Chen, Dana Arad, and Gabriele Sarti (eds.), *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 109–136, Stroudsburg, PA, USA, 2025. Association for Computational Linguistics. ISBN 9798891763463. doi: 10.18653/v1/2025.blackboxnlp-1.6. URL http://dx.doi.org/10.18653/v1/2025.blackboxnlp-1.6.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak, and C Zhang (eds.), *Advances in Neural Information Processing Systems 37*, volume 37, pp. 124198–124235, San Diego, California, USA, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). doi: 10.52202/079017-3946. URL http://dx.doi.org/10.52202/079017-3946.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Jiaqi Mu and Pramod Viswanath. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of Preference Optimisation with DPO-Positive. *arXiv [cs.CL]*, 2024. doi: 10.48550/arXiv.2402.13228. URL http://arxiv.org/abs/2402.13228.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. Iterative Reasoning Preference Optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4XIKfvNYvx.

R L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975. URL https://academic.oup.com/jrsssc/article-abstract/24/2/193/6953554.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. Outlier Dimensions that Disrupt Transformers are Driven by Frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1286–1304, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.93. URL http://dx.doi.org/10.18653/v1/2022.findings-emnlp.93.

Qwen Team, An Yang, Baosong Yang, et al. Qwen2.5 Technical Report. *arXiv [cs.CL]*, 2024. doi: 10.48550/arXiv.2412.15115. URL http://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/pdf?id=HPuSIXJaa9.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak, and C Zhang (eds.), *Advances in Neural Information Processing Systems 37*, volume 37, pp. 126207–126242, San Diego, California, USA, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). doi: 10.52202/079017-4009. URL http://dx.doi.org/10.52202/079017-4009.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization. In *International Conference on Learning Representations*, volume 2025, pp. 24791–24834, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3df38ca67befaed9c03b95ffee07d9f8-Abstract-Conference.html.

Yi Ren and Danica J Sutherland. Learning Dynamics of LLM Finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tPNHOoZFl9.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL http://dx.doi.org/10.18653/v1/2024.acl-long.828.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference Ranking Optimization for human alignment. In *Proc. Conf. AAAI Artif. Intell.*, volume 38, pp. 18990–18998. Association for the Advancement of Artificial Intelligence (AAAI), 2024. doi: 10.1609/aaai.v38i17.29865. URL https://dl.acm.org/doi/10.1609/aaai.v38i17.29865.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In A Globerson, L Mackey, D Belgrave, A Fan, U Paquet, J Tomczak, and C Zhang (eds.), *Advances in Neural Information Processing Systems 37*, volume 37, pp. 125019–125049, San Diego, California, USA, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). doi: 10.52202/079017-3970. URL http://dx.doi.org/10.52202/079017-3970.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive Activations in Large Language Models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Team OLMo, Pete Walsh, Luca Soldaini, et al. 2 OLMo 2 Furious. *arXiv [cs.CL]*, 2024. doi: 10.48550/arXiv.2501.00656. URL http://arxiv.org/abs/2501.00656.

Team Olmo, Allyson Ettinger, Amanda Bertsch, et al. Olmo 3. *arXiv [cs.CL]*, 2025. doi: 10.48550/arXiv.2512.13961. URL http://arxiv.org/abs/2512.13961.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, et al. LLaMA: Open and efficient foundation language models. *arXiv [cs.CL]*, 2023a. URL http://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*, 2023b. doi: 10.48550/arXiv.2307.09288. URL http://arxiv.org/abs/2307.09288.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL http://dx.doi.org/10.18653/v1/2024.findings-emnlp.620.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural Text Generation With Unlikelihood Training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

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

Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A Common Pitfall of Margin-based Language Model Alignment: Gradient Entanglement. In *International Conference on Learning Representations*, 2025. arXiv:2410.13828.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-Rewarding Language Models. In *International Conference on Machine Learning*, pp. 57905–57923. PMLR, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv [cs.CL]*, 2024. doi: 10.48550/arXiv.2401.02385. URL http://arxiv.org/abs/2401.02385.

Appendix Overview

Appendix A extends the related work of §2. Appendix B collects the setup details, full results, and supporting measurements behind the main text, and ends with the formal statement of the common-shift invariance. The third column names the main-text section that each part supports.

Part	Contents	Supports	Page
A	*Extended related work.* Iterated preference optimization, likelihood displacement, likelihood regularizers, and representation geometry.	§2	15
B.1	*Full cross-family onset table.* All 17 lineages, the budget at which each collapses if it does, and the collapse criterion.	§4	16
B.2	*Setup and measurement details.* β per scene type, exposure, and the full definitions of every metric.	§3	16
B.3	*Training curves and extended-budget trajectories.* Flagship Llama-2 runs and six-fold-budget OLMo runs.	§4, §7	17
B.4	*Routes of the released mass at extended budget.* Where the mass goes under each penalty, and the residual penalty.	§7	17
B.5	*Penalty trajectories across families.* Offset norm and its largest vocabulary entry under each penalty.	§7	17
B.6	*A listwise advantage orthogonal to collapse.* Ordering of unseen rejected responses by objective.	§3	18
B.7	*Removing the offset from a collapsed model.* Subtracting its logit bias at decoding time undoes the collapse.	§5	18
B.8	*Absolute logit scale.* A z-loss compresses $\log Z$ without removing the well.	§5	19
B.9	*Dose along training and the objective's demand.* Dose per lineage, the sweep over β, and standard-recipe offsets.	§6	21
B.10	*Full-parameter onset.* Onset tracks the integrated learning rate.	§4, §5	21
B.11	*A supervised fine-tuning control.* Plain fine-tuning builds a shared offset but no well.	§6	22
B.12	*SimPO from a supervised fine-tuning checkpoint.* The official recipe from a public SFT model, one-fold and three-fold budgets.	§6	22
B.13	*External reward models.* Judge protocol, win rates of all runs, and ranking against the margin.	§6, §8	23
B.14	*Shapes of well that standard rulers miss.* Latent wells, non-repeating wells, and word-family rings.	§3, §8	24
B.15	*Supplementary measurements.* Further numbers behind the claims of the main text.	§5–§8	25
B.16	*Seeds and scope.* Seed repeats, adapter ranks, and the ranges of the demand analysis.	§9	26
B.17	*Common-shift invariance and the trained offset.* Two propositions and the measured split of the offset.	§3, §5	26

A Extended Related Work

Iterated preference optimization. Recent pipelines repeat preference optimization over several rounds on continually refreshed preference data (Yuan et al., 2024; Grattafiori et al., 2024), although sampling more responses per prompt does not always help (Xiao et al., 2025).

Likelihood displacement, squeezing, and gradient entanglement. Recent studies have documented the probabilistic symptoms of preference optimization, but from different angles. Razin et al. (2025) demonstrate that DPO training can decrease the chosen response's absolute likelihood. They trace this to the unembedding matrix's geometry, using a centered hidden-embedding similarity score to predict *when* a dataset risks spilling rejected updates onto chosen ones. Yuan et al. (2025) prove that margin-based losses entangle the gradients of chosen and rejected log-probabilities, explaining *why* the two tend to move in tandem. While these works map the symptoms in probability space, we isolate their source in representation space, the *common-mode offset* in the final hidden state. Furthermore, Ren & Sutherland (2025) show that the negative gradient on rejected tokens "squeezes" probability mass toward tokens the base model is already confident about. Pushing down the probability of unwanted tokens is also how unlikelihood training reduces repetition (Welleck et al., 2020), whereas in preference optimization the push on rejected responses can end in repetition. The vocabulary wells we discover represent a different phenomenon, since the base

model initially ranks our well tokens far below its preferred tokens, typically by 17 to 39 nat. It is the accumulation of the common-mode offset that inflates the logits of these low-confidence tokens, until one of them overcomes this gap and dominates decoding at most positions.

Regularizing absolute likelihood. Several methods anchor the absolute probabilities that a relative ranking loss leaves free. DPO-Positive (Pal et al., 2024) penalizes the chosen log-probability if it falls below its reference value; Pang et al. (2024) inject an explicit NLL term for reasoning tasks; and Im et al. (2026) regularize the length-normalized log-probability of the full response set toward the reference. However, because these interventions constrain the scalar probabilities rather than the underlying vector representations, they do not directly constrain the shared direction in the hidden state. In our experiments, an NLL term on the chosen response delays the growth of the common-mode offset without eliminating it (Figure 5). When it does prevent immediate collapse, the vocabulary well relocates to positions *after* the desired answer, where the NLL term exerts no constraint.

Unembedding geometry and representation anisotropy. Prior work has mapped the geometry of pretrained representations. Word embeddings and contextual hidden states are anisotropic, sharing a common mean and a few dominant directions (Mu & Viswanath, 2018; Ethayarajh, 2019; Biś et al., 2021), and standard training often squeezes them into a narrow cone (Gao et al., 2019). Furthermore, specific outlier dimensions are known to track token frequency (Puccetti et al., 2022; Macocco et al., 2025), encode massive activations (Sun et al., 2024), or regulate output entropy (Stolfo et al., 2024). However, these works describe the geometry of a *pretrained* model. The common-mode offset we study is constructed during preference *training*. It does not reshape the entire embedding space into a cone; instead, it acts as a single, shared displacement vector added to the hidden states, which increases the inner product with one specific vocabulary row. Functionally, this resembles activation steering, where adding a fixed vector at every position predictably alters a model's behavior (Rimsky et al., 2024). Mechanistic studies of DPO also find that it changes behavior by shifting the residual stream, bypassing rather than removing capabilities the model already has (Lee et al., 2024). The key distinction is that preference optimization grows this steering vector on its own in the background, and the behavior it ultimately produces is generation collapse. Finally, Land & Bartolo (2024) identify "under-trained" tokens with near-zero frequencies in base models. While many of the severe vocabulary wells in our experiments happen to emerge from this vulnerable population, these fragile tokens do not spontaneously cause collapse; they merely provide a path of least resistance for the common-mode offset. Decoding-time corrections that subtract unconditional logit terms (Ding et al., 2026) act after the fact and leave the training-time cause in place.

B Additional Details

B.1 Full Cross-Family Onset Table

Table 4 reports every base-model lineage with at least one untreated run.

B.2 Setup and Measurement Details

On the driving lists, β is fixed per scene type, 0.12 for the two thirds of lists in normal driving and 0.18 to 0.40 for the other scene types; on UltraFeedback, $\beta = 0.12$. SimPO's target margin γ is subtracted from the top response's scaled score at each stage. Because a list of length K yields $K - 1$ comparisons, the response ranked k-th is rejected in $k - 1$ of them, a quantity we call its *exposure*. In the flagship run the rejected responses fall in order of their exposure, so the more often a response serves as a negative baseline, the steeper its drop.

We evaluate each trained run along four axes, since no single metric captures the whole failure. *Mass flow* is defined in §3. *Degeneration* combines standard metrics for n-gram repetition (Holtzman et al., 2020; Welleck et al., 2020), non-ASCII runs, and abnormal length with two measures of our own. A word-stem ring rate catches token-family wells, which we call *word-family rings*, and the base model's per-token negative log-likelihood of the samples catches degeneration that does not repeat. *Held-out ranking* uses 600 unseen lists of a prompt and its responses, or 300 of them for the Mistral runs, the six-fold OLMo-2 runs, and the training trajectories. Top-1 accuracy is the fraction of lists whose chosen response scores highest, and pairwise accuracy is the fraction of response pairs within a list that the model orders correctly. Kendall's τ (Kendall, 1938) is computed over the

Table 4: Fourteen of the 17 base-model lineages collapse within six times the base budget (§4). A run collapses once one token takes most of the released mass, a well, or once most of its samples degenerate or ring on one word stem, a word ring. Each row shows the untreated run at the onset budget, or at the largest budget tested when no collapse occurs. The onset multiple is relative to the 504-step base budget under the fixed schedule. TinyLlama, the three smallest Qwen2.5 models, and Llama-3.2-3B were trained at one and six times the budget only, Qwen2.5-14B first at three times, Phi-3-mini and Gemma-2-2B at one and three times, and Gemma-2-9B at the base budget only. Ring rates marked – come from audits that predate the ring measure, and † marks a count on the 12 saved samples. Each family is cited at its first row.

Base	Vocabulary	Onset multiple	Capture	Degen. / ring	Collapse
Llama-2-7B (Touvron et al., 2023b)	32k	1×	89.4%	96.0% / –	well
LLaMA-1-7B (Touvron et al., 2023a)	32k	1×	53.0%	86.3% / –	well
TinyLlama-1.1B (Zhang et al., 2024)	32k	6×	99.0%	100% / 100%	well
Llama-3.2-1B-Instruct (Grattafiori et al., 2024)	128k	1×	88.8%	61.0% / –	well
Llama-3.2-3B	128k	none within 6×	43.7%	10.7% / 7.7%	no
Llama-3.1-8B	128k	1×	87.6%	18.3% / 22.0%	well
Qwen2.5-0.5B (Qwen Team et al., 2024)	152k	6×	93.9%	7.7% / 6.7%	well
Qwen2.5-1.5B	152k	6×	93.8%	99.7% / 1.3%	well
Qwen2.5-3B	152k	6×	89.6%	91.3% / 0.3%	well
Qwen2.5-7B	152k	3×	93.9%	99.3% / –	well
Qwen2.5-14B	152k	3×	27.0%	80.3% / 78.7%	word ring
Mistral-7B-v0.3 (Jiang et al., 2023)	32k	1×	95.1%	100% / 0%	well
OLMo-2-7B (Team OLMo et al., 2024)	100k	6×	89.6%	21.0% / 8.0%	well
OLMo-3-7B (Team Olmo et al., 2025)	100k	3×	93.8%	96.0% / 0%	well
Gemma-2-2B (Gemma Team et al., 2024)	256k	none within 3×	0.4%	4.7% / 4.3%	no
Gemma-2-9B	256k	1×	12.5%	0% / 11 of 12†	word ring
Phi-3-mini (Abdin et al., 2024)	32k	none within 3×	18.4%	15.7% / 38.7%	no

unseen rejected responses, and the margin is the chosen response's mean log-probability minus that of the highest-ranked rejected response. *Offset geometry* covers the norm $\|\bar{b}\|$, the loudness, and, for candidate well tokens, their unembedding row-norm percentile and their counts in the training text and in the 10^8-token WikiText-103 corpus (Merity et al., 2017). For centering, $\bar{b}$ is computed on the current batch against the adapter-free forward pass. Collapse and ranking dynamics are measured across all lineages, while the computationally intensive analyses of §5 and §6 use the representative subsets named there.

B.3 Training Curves and Extended-Budget Trajectories

The runs without offset control drive the ranking loss lowest and the training margin highest, and almost all of their extra margin comes from pushing the rejected responses down (Figure 4). Figure 5 shows the six-fold-budget OLMo trajectories behind §7, including the per-position loudness penalty runs.

B.4 Routes of the Released Mass at Extended Budget

At six times the base budget, each penalty on the OLMo family closes one route for the released mass, and the mass takes the next one (Figure 5). Penalizing loudness alone moves the well to another token of the same token family. Adding the norm penalty removes the well but sends about half of the released mass to punctuation and premature end-of-text tokens. Applying the loudness penalty at every position, rather than only to the mean offset, does not close this route either. Combined with the norm penalty, it still leaves about 40% of the mass on punctuation and end-of-text on OLMo-3, and without the norm penalty it turns most OLMo-2 samples into word-family rings. Penalizing the residual itself, with $\lambda_{\text{res}} \frac{1}{T} \sum_t \|r_t\|$, closes both routes on OLMo-3 and leaves no well, no tail, and unchanged held-out ranking. On OLMo-2, however, the mass moves on into word-family rings and the deep tail.

B.5 Penalty Trajectories Across Families

Figure 6 shows how each penalty shapes the offset over training on Llama-2 and Qwen. The norm penalty keeps the offset short, while the loudness penalty lets it grow but keeps its largest vocabulary entry small. On Qwen, the loudness penalty alone lets the offset grow to 490 to 550, more than ten

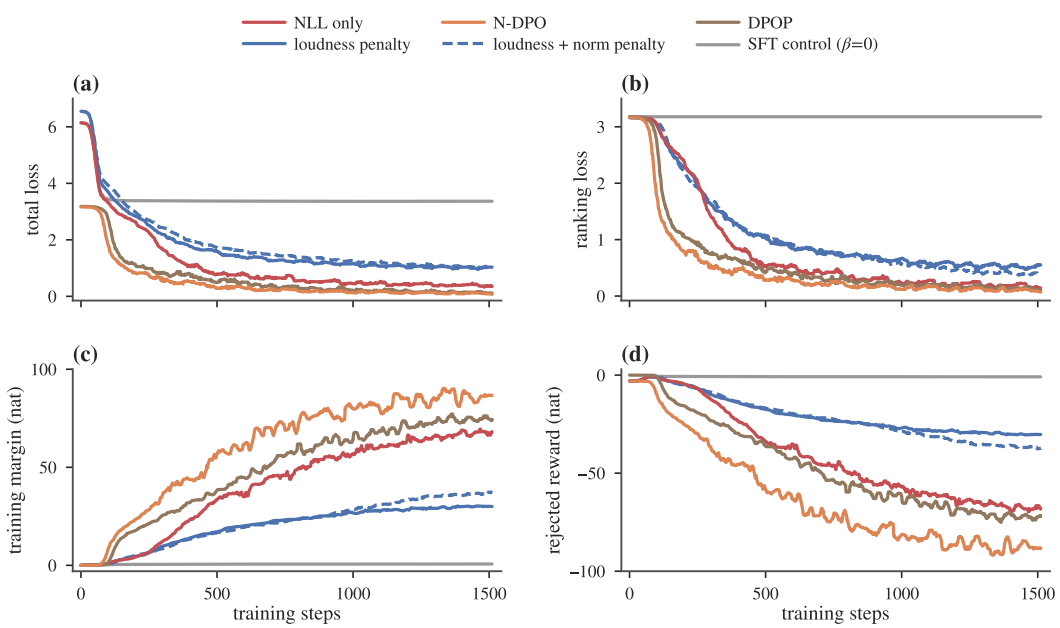

Figure 4: Training curves of six Llama-2-7B runs with $K{=}4$ sequential PL on the driving lists at three times the base budget, smoothed with a 25-step running median. (a) Total loss and (b) ranking loss fall lowest in the three runs without offset control, NLL only, N-DPO, and DPO-Positive (DPOP), which end with 71 to 81% of the released mass in a single well. The ranking loss of the SFT control ($\beta{=}0$) stays at its initial value. (c) Without offset control the training margin reaches 67 to 87 nat, against 30 and 37 nat with it, and (d) almost all of this difference comes from pushing the rejected responses down.

times its length under the norm penalty, which is consistent with the family-specific penalty strengths of §7.

B.6 A Listwise Advantage Orthogonal to Collapse

Held-out evaluation reveals one respect in which listwise and pairwise objectives differ, independent of collapse. Beyond whether the single best response is identified, we ask whether a model preserves the graded ordering among the *unseen rejected* responses of a held-out list. Pairwise training never supervises this ordering, since it only compares the top response against one alternative. On the driving lists, pairwise ranking ($K{=}2$) inverts the graded ordering, with ordering accuracy of 0.17 to 0.27 and Kendall τ of -0.46 to -0.66, while every sequential listwise $K{=}4$ run preserves it, at 0.82 to 0.92 (Figure 7). All-pairs and $K{=}3$ training fall in between at about 0.58, and SFT alone falls slightly below chance, at 0.415. On UltraFeedback the effect does not appear, and pairwise training instead slightly improves the ordering of rejected responses over the base model. This listwise benefit is therefore domain-dependent and orthogonal to the collapse mechanism of §5. It does not correlate with whether a run nucleates a well, and offset-controlled runs keep whichever ordering their underlying objective produces.

B.7 Removing the Offset from a Collapsed Model

We subtract a multiple of the logit bias $W\bar{b}$ from the untreated flagship model at decoding time, with $\bar{b}$ measured on 80 rejected responses, and decode 60 held-out prompts greedily for 100 tokens. Subtracting $\bar{b}$ from the final hidden state gives the same result, since the unembedding is linear. Removing half of the bias stops the collapse, and removing all of it works nearly as well (Table 5). Removing only the well's own logit is not enough, because the collapse then moves to the next token of the same direction, a rare Cyrillic token. The recovered model continues text normally but does

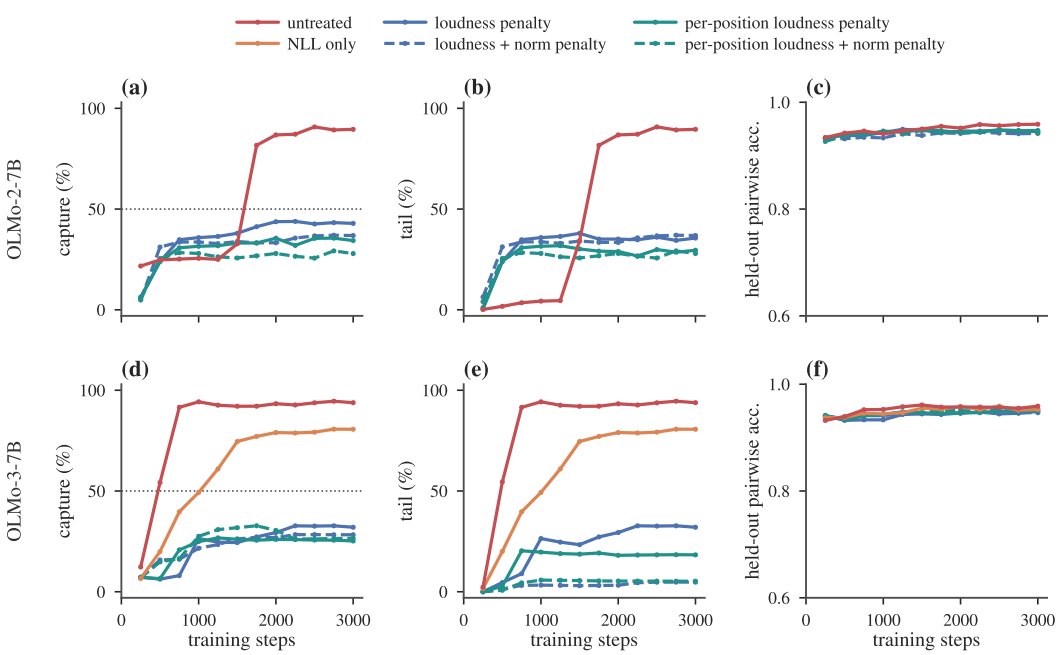

Figure 5: Six-fold-budget training trajectories of OLMo-2-7B (a–c) and OLMo-3-7B (d–f) in six conditions, namely untreated, NLL only, and four offset-control variants that penalize loudness, loudness and norm, per-position loudness, or per-position loudness and norm. OLMo-2 has no NLL-only run. (a) and (d) Well capture crosses the dotted 50% line in the untreated and NLL-only runs and ends at 81 to 94%, while no offset-control variant exceeds 45%. (b) and (e) The tail beyond rank 10,000 rises with the capture in those runs. With offset control it ends at 28 to 37% on OLMo-2, but near 5% on OLMo-3 once the norm penalty is added. (c) and (f) Pairwise accuracy on 300 held-out lists stays between 0.93 and 0.96 in every run.

not produce the answer format, which the untreated run never learned, so removing the offset stops the damage without aligning the model.

Table 5: Subtracting the offset's logit bias from the collapsed flagship model at decoding time. Degen. is the share of degenerate greedy samples, and the log-probabilities are per-token means under teacher forcing with the same subtraction.

Subtracted	Degen.	Chosen log-prob (nat)	Rejected log-prob (nat)
Nothing	100%	−13.37	−65.01
Only the well's logit	22%	−8.73	−54.64
The eight largest entries of $W\bar{b}$	65%	−7.17	−48.76
Half of $W\bar{b}$	0%	−4.70	−39.44
All of $W\bar{b}$	7%	−5.34	−24.98

B.8 Absolute Logit Scale

One alternative hypothesis is that collapse is a symptom of the output distribution's absolute scale drifting upward, since a softmax is unchanged when the same constant is added to every logit. We test it by penalizing $\log Z = \log \sum_v \exp(z_v)$, a z-loss as in Chowdhery et al. (2023), at increasing strength on the Llama-2 flagship setting (Figure 8). The strongest coefficient compresses $\log Z$ 2.5-fold, shrinks the relative peak logit and the common-mode vocabulary projection, and lowers the margin by about one third relative to the untreated run. Held-out ranking accuracy is essentially unchanged, and well capture stays at 65 to 66% of the released mass, against 74% untreated. At this coefficient the z-loss gradient reaches 19 to 43% of the ranking-gradient norm, yet collapse persists. We treat this as supporting rather than conclusive evidence, because a raw gradient-norm ratio is not the same as its realized contribution to the optimizer update. Together with §5, the result

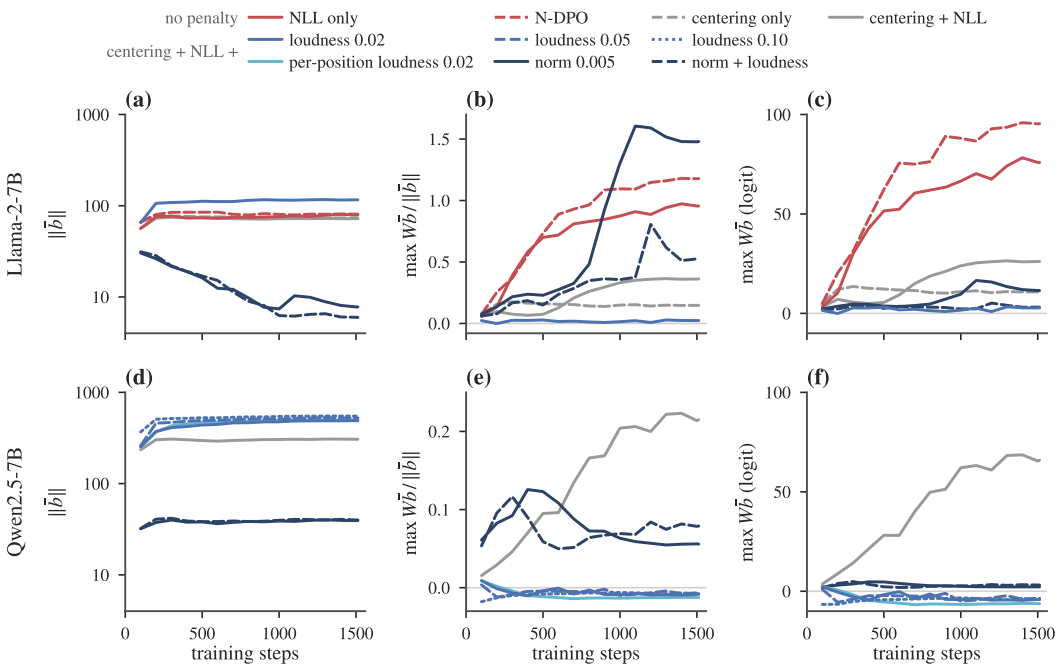

Figure 6: (a) Across seven Llama-2-7B runs, $\|\bar{b}\|$ stays between 70 and 85 without a penalty, grows to 116 under the loudness penalty, and falls to 6 to 8 once the norm penalty is added. (b) Divided by $\|\bar{b}\|$, the largest entry of $W\bar{b}$ stays near zero only under the loudness penalty. (c) NLL only and N-DPO lift the largest entry of $W\bar{b}$ to 76 and 95 by the end of training, while the loudness penalty holds it below 4. (d) Across seven Qwen2.5-7B runs, $\|\bar{b}\|$ grows to 490 to 550 under every loudness penalty but stays near 40 once the norm penalty is added. (e) Divided by $\|\bar{b}\|$, the largest entry of $W\bar{b}$ falls below zero under every loudness penalty. (f) Without a penalty, the largest entry of $W\bar{b}$ climbs to 66, while every penalized run stays below 5.

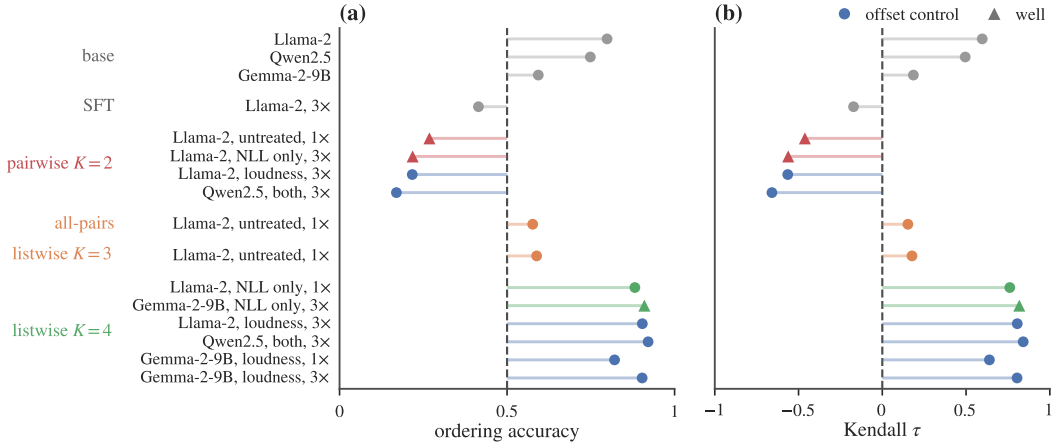

Figure 7: (a) On 600 held-out lists, pairwise $K{=}2$ training orders the unseen rejected responses below chance (dashed line), with or without offset control (blue), while every sequential listwise $K{=}4$ run orders them better than the base models. (b) The Kendall τ over the rejected responses alone is negative for SFT and every pairwise run, and at least 0.64 for every listwise $K{=}4$ run. Triangles mark runs that form a well.

indicates that the operative quantity is not the overall output scale but the offset's projection onto the vocabulary, which a z-loss constrains only indirectly and incompletely.

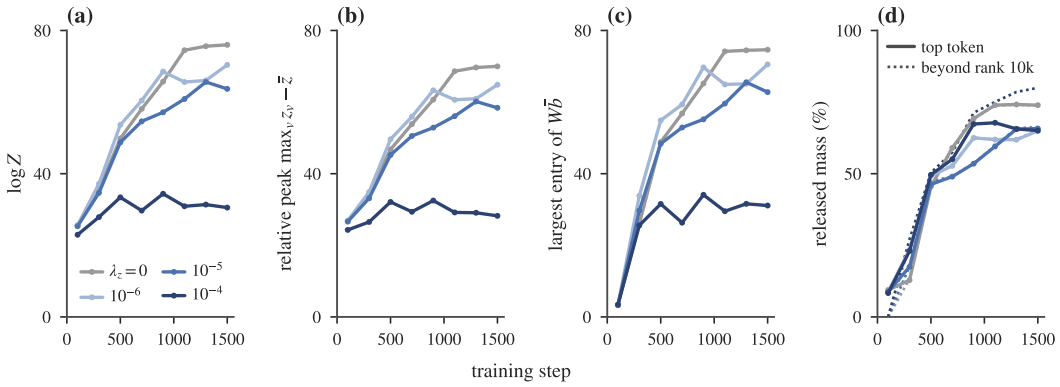

Figure 8: A z-loss penalty on $\log Z$ compresses absolute logit scale, relative peak, and common-mode vocabulary projection without preventing well capture. Blue lines show three strengths of λ_z on the Llama-2 flagship setting, darker for stronger, and the gray line shows the untreated run. (a) The strongest penalty, $\lambda_z = 10^{-4}$, lowers $\log Z$ from 76.0 to 30.5 by step 1500. (b) It lowers the relative peak logit by the same factor of 2.5, from 70.0 to 28.2. (c) It lowers the largest entry of $W\bar{b}$ from 74.6 to 31.1. (d) The top token still takes 65 to 66% of the released mass at every strength, against 74% untreated, and at $\lambda_z = 10^{-4}$ the mass beyond rank 10,000 (dotted) rises to 80%.

B.9 Dose Along Training and the Objective's Demand

We summarize each snapshot by its median dose, the median gain divided by the median base gap over the rejected positions. In all nine lineages traced every 50 steps, the median dose passes one before the well takes over, and the takeover itself takes 100 to 300 steps (Figure 9). Across the sweep over β, the final dose passes one only under per-token averaging with a demand above the base gap (Figure 10). The standard recipes build the same kind of offset as the collapsing runs, at a smaller dose (Figure 11).

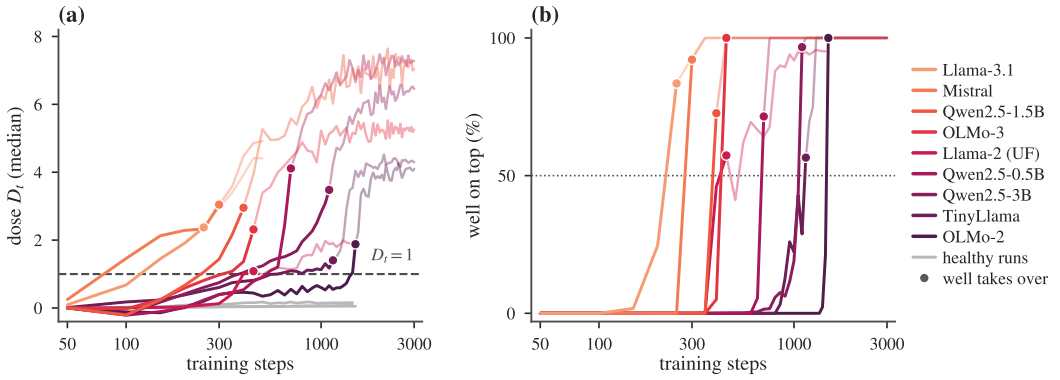

Figure 9: Dose along natural training runs, with a dot at the first snapshot at which the well is the top token at a majority of rejected positions. (a) In all nine collapsing lineages, the median dose $D_t = g_t/m_t$, measured every 50 steps, passes one before the dot and lies between 1.1 and 4.1 at the dot, while the three runs that stay healthy, in gray, remain below 0.2. (b) The share of rejected positions with the well on top climbs from under 10% to a majority within 100 steps in seven lineages, and within 200 and 300 steps in Qwen2.5-3B and TinyLlama.

B.10 Full-Parameter Onset

Under full-parameter fine-tuning, onset tracks the integrated learning rate, the learning rate summed over the steps taken. Two Llama-2 runs, at learning rates 5×10^{-7} and 10^{-6} under a 1512-step schedule, nucleated once it reached about 1.7×10^{-4}. Before running it, we then wrote down that a

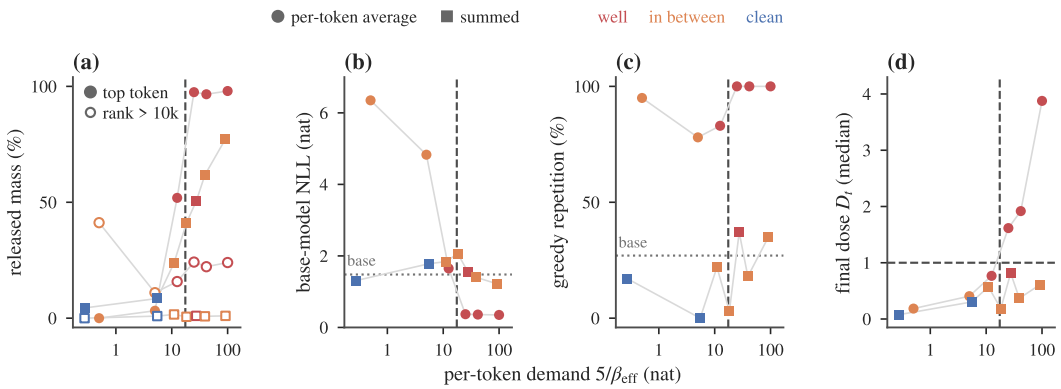

Figure 10: Outcome of pairwise training on UltraFeedback against the per-token margin the objective demands, $5/\beta_{\text{eff}}$ with $\beta_{\text{eff}} = \beta$ for per-token-average scores and $\beta|y|$ for summed scores. The dashed vertical line marks the median base gap at the rejected positions, 17.5 nat, and dotted lines mark the base model at 1.48 nat and 27%. (a) Under per-token averaging the well takes more than 95% of the released mass once the demand passes the base gap, whereas summed scores concentrate mass gradually and send almost none beyond rank 10,000. (b) Samples from the two lowest-demand per-token runs have a high negative log-likelihood under the base model, samples from per-token runs beyond the base gap a low one, and summed runs stay near the base model. (c) Every per-token run repeats under greedy decoding on at least 78% of prompts, against at most 37% for summed runs. (d) The final dose passes one only under per-token averaging with a demand above the base gap. Red marks runs with more than half of the released mass on one token and at least 30% degenerate samples, blue marks clean runs, and orange the runs in between. Doses are for each run's leading token, or for `Conf` when that token takes less than 5% of the released mass.

run at 10^{-6} under the 504-step schedule would nucleate between steps 200 and 300, and it nucleated at step 250. In all full-parameter runs with the per-token objective, the offset alone makes the well the top token at 98 to 100% of positions, while the fine-tuned unembedding rows alone do so at no more than 1.5%.

B.11 A Supervised Fine-Tuning Control

To check whether the offset needs a ranking term at all, we trained Llama-2-7B on the driving lists at three-fold budget with only the NLL term on the chosen response. Its offset reaches a norm of 56, and the offsets on chosen and rejected responses have a cosine of 0.99. However, the largest entry of $W\bar{b}$ is only 3.4 and the top token takes 8.6% of the released mass, so no well forms. Plain fine-tuning thus already builds a shared offset, while the ranking term decides where it points.

B.12 SimPO from a Supervised Fine-Tuning Checkpoint

We start from princeton-nlp/Llama-3-Base-8B-SFT, the Llama-3-8B model that the SimPO authors fine-tuned on UltraChat (Meng et al., 2024), and train it with their Llama-3-Base configuration of $\beta = 2.0$, $\gamma = 1.0$, a learning rate of 6×10^{-7}, a global batch of 128, a cosine schedule with a warmup ratio of 0.1, and full-parameter updates. The data are our 15.7k UltraFeedback pairs rendered in the model's chat template. The base budget of 369 steps makes three passes over these pairs, as many examples as about three quarters of one epoch of the official training data, and the three-fold budget of 1107 steps makes nine. We keep fp32 master weights with an 8-bit AdamW optimizer, where the authors use bf16 weights and AdamW.

Before training, we measured the base gap of three candidate wells at the rejected positions. Relative to Llama-3-8B, the SFT checkpoint enlarges the median base gap by 22% for `Conf` and by 10% for `Apache` and shrinks it by 2% for `Sharper`. Transplanting the offset of the Llama-3.1 full-parameter SimPO run at three-fold budget, whose well is `Apache`, gives a median dose of 1.12 on the SFT checkpoint against 1.27 on Llama-3-8B. By the dose, supervised fine-tuning should therefore delay a well rather than prevent it.

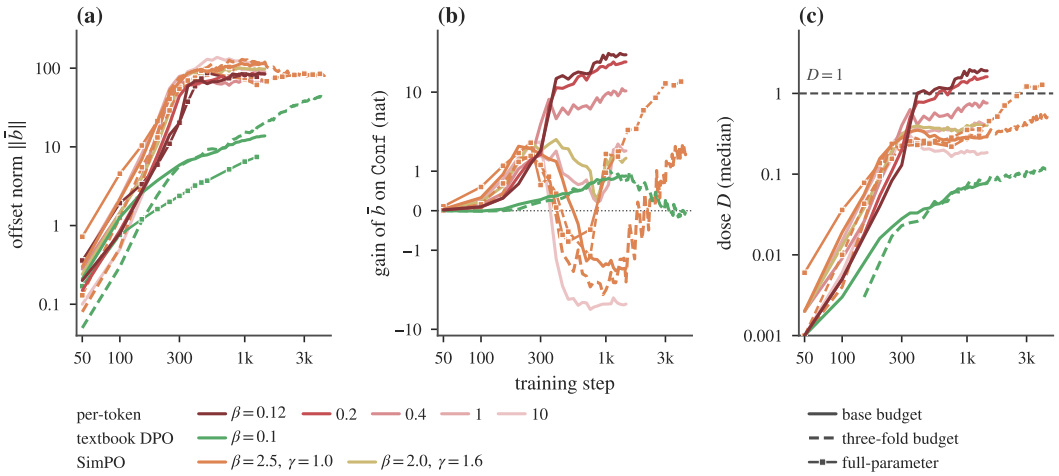

Figure 11: Standard recipes build the same offset at a smaller dose. All runs train Llama-2-7B pairwise on UltraFeedback and are plotted against training step on log axes. Per-token averaging at $\beta \leq 0.2$ collapses, while textbook DPO at $\beta = 0.1$ and SimPO at $\beta \in \{2.0, 2.5\}$ do not within the base budget. (a) By the end of the base budget, SimPO has built an offset about as long as that of the collapsing runs, and textbook DPO one at most a sixth as long. (b) Only per-token averaging at $\beta \leq 0.2$ gives `Conf`, the well of the collapsing run, a logit gain above 20 nat, while textbook DPO and SimPO stay below 3 nat within the base budget. (c) By the end of the base budget, the doses of textbook DPO and SimPO sit four to twenty-five times below those of the collapsing runs, and full-parameter SimPO crosses $D = 1$ at step 3000 under the three-fold budget. Squares mark full-parameter fine-tuning, and dashed lines mark three-fold-budget runs, which train for nine epochs under their own schedule. Full-parameter per-token averaging, which collapses onto the space token, and full-parameter textbook DPO appear only in (a).

Table 6 reports both budgets. No well forms, and held-out ranking improves over the SFT checkpoint at both budgets. From one-fold to three-fold budget, the offset grows from 32.5 to 77.3 while its largest vocabulary entry stays small. At the base budget it lifts the ordinary `Conf` family by about 2 nat, and at three times the budget a flat spread of infrequent multilingual and code fragments by about 4 nat. SimPO from the base models lifts its leading tokens by about 18 nat. No token in the vocabulary reaches a median dose of 0.4, while SimPO from the base models carries `Apache` on Llama-3.1 and `Conf` on Llama-2 past one. Meanwhile the samples move away from the SFT model, whose per-token negative log-likelihood of them rises from 1.35 to 3.86. Against the SFT model's own samples for the same 300 held-out prompts, Skywork-Reward-V2 and ArmoRM prefer the one-fold run's samples on 63% and 65% of prompts and the three-fold run's on 34% and 33%. The held-out chosen responses win on 73% and 72% (Table 7).

Table 6: SimPO from the public SFT checkpoint with the official hyperparameters. NLL is the SFT model's per-token negative log-likelihood of sampled responses, the dose is for `Conf`, and held-out pairwise accuracy is 0.619 for the SFT checkpoint itself.

Budget	Steps	Top token	Capture / tail	Degen.	NLL	$\|\bar{b}\|$	Dose	Held-out top-1 / pairwise
1×	369	,	2.5% / 0.6%	1.3%	1.35	32.5	0.15	0.41 / 0.67
3×	1107	(	2.7% / 4.1%	0.0%	3.86	77.3	0.19	0.58 / 0.75

B.13 External Reward Models

Table 7 lists the reward-model evaluation behind §8 and the Win column of Table 2. Each run generates one sample for each of the same 300 held-out UltraFeedback prompts, at temperature 0.7 with nucleus sampling at 0.95 and at most 256 new tokens. Each sample is compared with the sample of the model the run starts from for the same prompt, which is the base model except for the two runs from the SFT checkpoint. The chosen responses of the held-out lists win against the base

model on 96% and 95% of prompts under the two reward models. Adding the SimPO offset alone to the base model, as in §5, keeps its samples about as likely under the base model as the base model's own samples. They cost about 0.6 nat per token, against 3.1 to 4.3 for the trained SimPO models, and when they fail they fall into ordinary repetition loops rather than strings of unrelated words.

Table 7: External reward models on 300 held-out UltraFeedback prompts. Win is the share of prompts on which the reward model prefers the run's sample to that of the model the run starts from. Pairwise is held-out pairwise accuracy. Top-10 is the share of released mass that goes to the base model's ten most likely tokens under teacher forcing. The base Llama-2-7B has a held-out pairwise accuracy of 0.60, the base Llama-3.1-8B 0.61, and the SFT checkpoint 0.62. The last four Llama-2 rows train on four-response lists and the others on pairs.

Run	Skywork win	ArmoRM win	Pairwise	Top-10
Llama-2-7B				
Per-token ranking, $\beta = 0.12$	25%	6%	0.68	0.02
with a reference model, $1\times$ / $3\times$	26% / 17%	7% / 3%	0.68 / 0.71	0.04 / 0.05
full-parameter	24%	45%	0.70	0.14
SimPO, $1\times$ / $3\times$	7% / 6%	3% / 4%	0.67 / 0.73	0.04 / 0.04
SimPO, $\beta = 2.0$, $\gamma = 1.6$	8%	6%	0.67	0.05
SimPO, full-parameter, $1\times$ / $3\times$	10% / 7%	6% / 4%	0.70 / 0.76	0.10 / 0.05
Textbook DPO, $1\times$ / $3\times$	68% / 66%	67% / 65%	0.63 / 0.64	0.77 / 0.77
Textbook DPO, full-parameter	66%	67%	0.63	0.80
NLL only, $\lambda = 1$	64%	64%	0.63	0.75
Centering only	20%	23%	0.68	0.42
Recipe, $\lambda_{\text{NLL}} = 1.0$ / 0.1	65% / 52%	66% / 55%	0.63 / 0.66	0.74 / 0.45
Centering and loudness penalty	17%	15%	0.69	0.17
Llama-3.1-8B				
Per-token ranking, $\beta = 0.12$	11%	5%	0.74	0.02
SimPO, full-parameter, $1\times$ / $3\times$	13% / 3%	14% / 5%	0.78 / 0.79	0.01 / 0.03
Llama-3-8B from the SimPO authors' SFT checkpoint				
SimPO, official recipe, $1\times$ / $3\times$	63% / 34%	65% / 33%	0.67 / 0.75	0.67 / 0.46

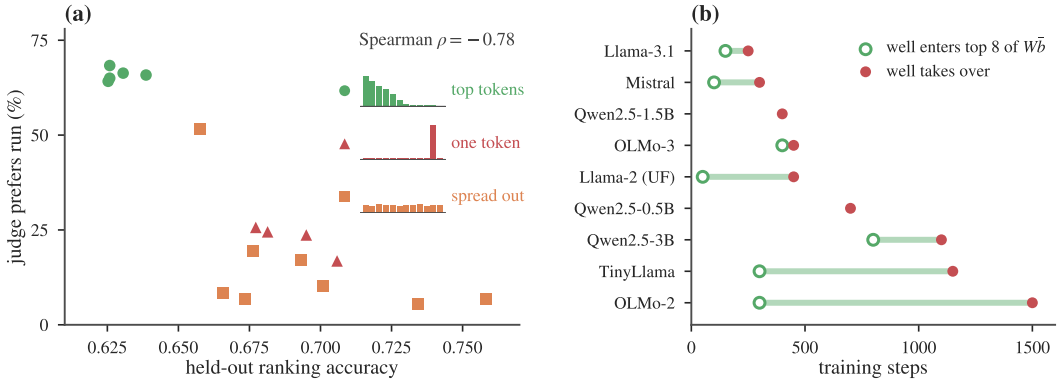

Figure 12: (a) Across 17 UltraFeedback runs, the better a run ranks held-out responses, the less often a public reward model prefers its samples to the base model's; the glyphs show where each group's released mass goes. (b) In seven of nine collapsing lineages, the eventual well enters the eight largest entries of $W\bar{b}$ before it takes over the rejected positions.

B.14 Shapes of Well That Standard Rulers Miss

We compared the standard n-gram repetition ruler with our mass-flow and word-stem measurements on every run. The standard ruler correctly flags the flagship run's repetition loop, in which 96% of samples degenerate, but it misses three other shapes of well. A *latent well* appears only after the visible answer ends, meaning greedy decoding within a length-matched budget never reaches it. A *non-repeating well* keeps changing its surface form, for example a code-path continuation,

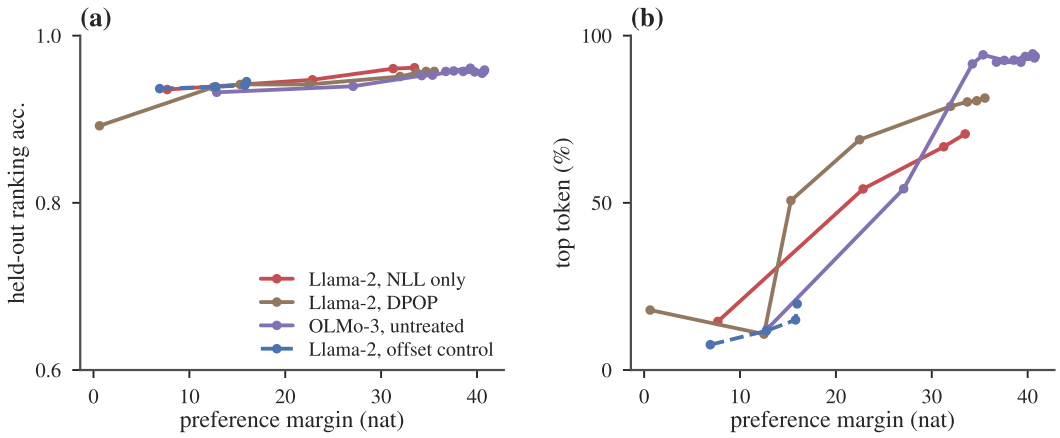

Figure 13: (a) Held-out ranking accuracy and (b) the top token's share of released mass against the preference margin along training. Ranking saturates early for every run, while the top token's share keeps growing with the margin only in the runs without offset control.

while still capturing most of the released mass. The ruler then reports little repetition despite severe collapse. A *word-family ring* lacks a single dominant token, because a cluster of low-frequency, morphologically related tokens takes the released mass together. A ruler that counts single tokens therefore underreports the failure. Furthermore, these standard rulers miss the diffuse degradation from §6, which leaves no repetitive patterns to count but is exposed by evaluating the base model's likelihood of the generated samples. Any conventional ruler used in isolation can therefore pass a model that the offset and mass-flow measurements from §3 reveal has already collapsed.

B.15 Supplementary Measurements

These measurements support claims of the main text, grouped by section.

The offset is shared (§5). Recomputed on held-out text with phrasing the model never saw, the flagship offset keeps a cosine of 0.85 with the original.

Dose threshold (§6). When the well takes over along natural training runs, it goes from under 10% of the rejected positions to a majority within 100 to 300 steps (Figure 9b).

Anomalous rows (§6). For every single-token well, we looked up the norm of its unembedding row and counted how often the token occurs in our training text and in WikiText-103. Most single-token wells of the adapter runs on the driving data never appear in our training text. In 13 bases, the highest-norm row belongs to a rare token, and its gain g_t can overtake base gaps of 17 to 39 nat. Under full-parameter fine-tuning on UltraFeedback, the Llama family collapses onto the bare space token, and there too the offset alone makes the well the top token at nearly every position.

Losses and demand (§6). On the driving lists, the per-token loss with a reference model collapses onto `textt` at the base budget and onto a rare Cyrillic token at the three-fold budget. Per-token training at β of 1 and 10, whose demand is also low, degrades in the same diffuse way (Figure 10). With three times the budget, the SimPO offset concentrates its projection on a few vocabulary rows even where it becomes shorter. By the end of full-parameter training, its leading token takes 27 to 45% of the released mass (Table 2).

Ranking cost of the recipe (§7). Eight base models have treated runs in our held-out evaluation. We compare each treated run with the untreated run of the same base and budget, or with the NLL-only run where no untreated run was scored. Every treated run but one lies between 1.8 points below and 0.1 points above its comparison run. The exception is the loudness variant of the recipe on Gemma-2-9B at the base budget, which barely lowers the rejected responses in training and scores 3.1 points below the untreated run, while the combined penalty on the same base scores 0.3 points below it.

Margin along training (§8). The margin reaches three or four times its value at the point where ranking saturates. At a matched margin of 12 to 16 nat, Llama-2 runs without offset control have

already channeled 42 to 51% of the released mass into a single well, against 8 to 20% for offset-controlled runs (Figure 13). The margin at which capture takes off also differs by base model, from 13 to 27 nat on six-fold OLMo-3 to 44 to 59 nat on OLMo-2, so the margin cannot serve as a stopping criterion.

B.16 Seeds and Scope

Both untreated seeds of the flagship settings collapse onto the same well as the first, `textt` on the driving lists and `Conf` on UltraFeedback. Their offset norms lie within 3% of the first seed's, and their dose crosses one no later than the well takes over. Both treated seeds stay free of a well, with 10 to 12% residual degeneration and unchanged held-out ranking. Adapter rank from 8 to 128 and attention-only adapters change when a well forms rather than whether. Across these ranks, offset norms lie between 80 and 117 on Llama-2 and Llama-3.1, and rank 8 on Llama-2 nucleates the same well at three-fold budget. The demand analysis of §6 covers pairwise UltraFeedback training under per-token averaging with β from 0.05 to 10, under textbook DPO with β from 0.1 to 0.0003, and under SimPO with β and γ from its published tuning range. These runs use our adapter and full-parameter protocols for three or nine epochs.

B.17 Common-Shift Invariance and the Trained Offset

Write $s_\theta(x, y)$ for the score of a response, the per-token average of §3 or the summed log-ratio of textbook DPO. Both $\mathcal{L}_{\text{PL}}$ and the DPO loss depend on the scores of a prompt's responses only through their differences, since the DPO reference term does not depend on θ.

Proposition 1 (Loss-level invariance)**.** *Let ℓ be invariant under adding a constant to the scores of all responses of a prompt, and write a score perturbation as $\varepsilon = \bar{\varepsilon} + \tilde{\varepsilon}$ with $\bar{\varepsilon}(x)$ its mean over the prompt's responses. Then $\ell(s + \varepsilon) = \ell(s + \tilde{\varepsilon})$, so the common part $\bar{\varepsilon}$ cannot be identified from the objective with any amount of preference data, and $\sum_y \partial\ell/\partial s(x, y) = 0$ for every prompt.*

Proof. Score differences within a prompt cancel $\bar{\varepsilon}$. Differentiating $\ell(s + c\mathbf{1}) = \ell(s)$ in c at $c = 0$ gives the gradient identity. □

Proposition 2 (First-order constraint on a shared hidden shift)**.** *Let W be the unembedding with rows w_v, h_t the final hidden state at position t and $p_t = \mathrm{softmax}(Wh_t)$. Shifting every hidden state of every response by the same δ changes a per-token-average score by*

$$\begin{aligned} s(x, y; \delta) - s(x, y; 0) &= \frac{1}{|y|} \sum_t \Big[\langle w_{y_t}, \delta \rangle - \log \mathbb{E}_{v \sim p_t} e^{\langle w_v, \delta \rangle} \Big] \\ &= \langle g_y, \delta \rangle - \frac{1}{2|y|} \sum_t \mathrm{Var}_{v \sim p_t} \langle w_v, \delta \rangle + O(\|\delta\|^3), \end{aligned}$$

with $g_y = \frac{1}{|y|} \sum_t \big(w_{y_t} - \mathbb{E}_{v \sim p_t} w_v \big)$ the mean gradient of the score with respect to the hidden states. Hence $\nabla_\delta L|_{\delta=0} = \mathbb{E}_x \sum_y \frac{\partial \ell}{\partial s(x,y)} g_y$, which for a pairwise logistic loss equals $-\mathbb{E}_x[a(x)(g_{y_w} - g_{y_l})]$ with $a(x) = \beta\sigma(-\beta(\Delta_\theta - \Delta_{\text{ref}})) > 0$ and $\Delta_{\text{ref}} = 0$ without a reference model. To first order the objective constrains δ along this single direction, and on the orthogonal complement only the second-order term remains, which enters through the difference between the responses' mean per-position variances.

Proof. The first equality is the change of $\log \mathrm{softmax}$ under the logit shift $W\delta$, the second the cumulant expansion of $\log \mathbb{E}\, e^{\langle w_v, \delta \rangle}$, and the gradient follows from the chain rule with coefficients that sum to zero within each prompt by Proposition 1. □

Three remarks bound the statement. The constraint is first-order and local, since the softmax is nonlinear and minibatch gradients differ from the expectation, so the complement is weakly rather than exactly unconstrained. The shift δ is not a parameter but the image of parameter updates through Jacobians shared across positions and both responses. This sharing is how a component in the weakly constrained part arises at all, and gradient entanglement (Yuan et al., 2025) describes the

coupling that produces it. Any term that depends on absolute scores, such as an NLL term on the chosen response, breaks the invariance, which is one reason such terms delay displacement (§2).

Table 8 measures how much of the trained offset lies in the weakly constrained part. For each run we add the offset δ of §3 to the base model's final hidden states on 200 held-out lists. For each list we record the change of the chosen per-token score, a^+, and of the mean rejected per-token score, a^-, and split it into a common part $c = (a^+ + a^-)/2$ and a differential part $d = (a^+ - a^-)/2$. The table reports the means of a^+, a^-, and c over the lists and the mean of $|d|$, which exceeds half the difference of the mean scores because d changes sign from list to list. We report the same two means for three norm-matched random directions, and the cosine between δ and $\bar{u} = \frac{1}{N}\sum_x \left(g_{y^{(1)}} - \frac{1}{K-1}\sum_{j=2}^{K} g_{y^{(j)}}\right)$, the base model's chosen-minus-rejected score gradient with every list and every rejected response weighted equally. For the three per-token UltraFeedback runs, which pair the chosen response with a randomly drawn rejected one, $\bar{u}$ is the direction of Proposition 2 with equal weights, and at the base model those weights vary across lists by only 5 to 7% of their mean. For the driving run $\bar{u}$ approximates that direction, since the $K = 4$ loss weights the rejected responses by rank. In the four collapsing runs the common part is 20 to 530 times the differential part and the cosine is at most 0.16. The offset carries the collapse-producing component without accounting for all of the likelihood displacement. On the driving run it alone lowers the chosen score by 65 nat while training lowered it by 21, so the position-specific residual partly compensates its effect on the chosen response, consistent with §5. For the two standard-recipe runs both parts are small, below 0.02 nat under textbook DPO and 0.25 and 0.06 nat under SimPO, because their quiet offsets lift no token near the top.

Table 8: The trained offset δ added to the base model's final hidden states on 200 held-out lists. a^+ and a^- are the mean changes of the chosen and mean rejected per-token scores in nat, c the mean of $(a^+ + a^-)/2$ and $|d|$ the mean of $|a^+ - a^-|/2$ over lists, "random" the same two means for three norm-matched random directions, and cos the cosine between δ and $\bar{u}$, the base model's chosen-minus-rejected score gradient.

Base	Data, objective	$\|\delta\|$	a^+	a^-	c	$\|d\|$	random $\|c\|$ / $\|d\|$	cos
Llama-2-7B	UltraFeedback, per-token $\beta = 0.12$ (collapses)	84	-14.5	-15.2	-14.9	0.68	0.30 / 0.06	0.03
Llama-2-7B	driving, $K = 4$ (collapses)	93	-65.1	-65.2	-65.2	0.12	0.49 / 0.02	0.16
Qwen2.5-7B	UltraFeedback, per-token $\beta = 0.12$ (collapses)	282	-26.0	-27.1	-26.6	1.36	2.15 / 0.31	-0.02
Llama-3.1-8B	UltraFeedback, per-token $\beta = 0.12$ (collapses)	109	-29.2	-30.1	-29.6	0.94	0.54 / 0.10	0.05
Llama-2-7B	UltraFeedback, textbook DPO $\beta = 0.1$	14	-0.005	-0.013	-0.009	0.014	0.011 / 0.008	0.07
Llama-2-7B	UltraFeedback, SimPO $\beta = 2.5, \gamma = 1.0$	114	-0.23	-0.26	-0.25	0.06	0.57 / 0.09	0.01