

Spectral DPO: Mitigating Validation Collapse via Heavy-Tailed Regularization

Ibne Farabi Shihab^{*†1} and Sanjeda Akter^{*1} and Anuj Sharma²

¹Department of Computer Science, Iowa State University

²Department of Civil, Construction & Environmental Engineering, Iowa State University
ishihab@iastate.edu

Abstract

Direct Preference Optimization removes the explicit reward model of RLHF but inherits its central failure mode: the preference-reward proxy can keep improving while downstream reasoning, factuality, or safety degrades. We ask whether this reward overoptimization, which we label *validation collapse* as an operational name for the trajectory we measure, is visible in weight-space geometry. Drawing on Random Matrix Theory and the Heavy-Tailed Self-Regularisation hypothesis, we propose Spectral DPO, which augments the DPO objective with a stochastic penalty computed from a truncated top- k Lanczos estimate of the spectra of effective adapted projections. The penalty holds the fitted tail exponent near a heavy-tailed reference, caps scale-normalised growth of the largest singular direction, and matches the estimated spectrum to a heavy-tailed target. Across ten matched seeds, three model families, and five benchmarks spanning instruction following, dialogue, mathematical reasoning, factuality, and adversarial robustness, Spectral DPO gains 2.33–2.42 points over the strongest included baseline and sustains accuracy through ten epochs where vanilla DPO degrades. Controls and a post-hoc spectral repair rule out norm clipping, KL anchoring, and length inflation, and support the relevance of spectral geometry without identifying a single causal statistic. All runs use rank-64 LoRA, so the evidence concerns effective-spectrum control under rank-constrained adaptation.

1 Introduction

Aligning large language models with human preferences is now an essential ingredient in deploying them as assistants, agents, and decision-support systems. The dominant pipeline pairs supervised fine-tuning with a preference-optimisation step that adapts the model toward responses humans

rate as preferable (Ouyang et al., 2022; Bai et al., 2022b). Reinforcement learning from human feedback (Christiano et al., 2017; Stiennon et al., 2020) has been the standard implementation of the preference step, but it is operationally fragile and computationally heavy, which has motivated a series of direct methods that derive an implicit reward from preference data and avoid an explicit reward model. Direct Preference Optimisation (Rafailov et al., 2023) is the canonical instance of this family, and it has become the default preference-optimisation method in many open-source instruction-tuning recipes.

A property the family inherits from earlier reward-based work, however, is its vulnerability to proxy over-optimisation. Gao et al. (2023) define reward-model overoptimisation as improvement in a proxy reward accompanied by deterioration in a gold objective, and Rafailov et al. (2024) study analogous behaviour in direct-alignment algorithms. The DPO objective is purely behavioural: it asks only that the model assign higher likelihood to preferred than to dispreferred responses, with no explicit reference to whether the model has improved on the underlying tasks the preferences were collected for. A model that learns to produce verbose, confidently phrased outputs that score well on preference annotation, while losing accuracy on held-out reasoning, has not been broken in any visible way by the loss; it has done exactly what the loss asked. Empirically, the pattern shows up as a slow rise of validation preference reward paired with stagnant or declining downstream accuracy, getting steadily worse as preference optimisation runs longer. We refer to this measured trajectory as validation collapse. The term is an operational label for the signature we measure in this paper, not a distinct theoretical phenomenon and not a contribution in itself; our contribution is the spectral regulariser and the analysis of effective weight-space geometry during the divergence.

^{*}Equal contribution.

[†]Corresponding author: ishihab@iastate.edu.

The natural place to look for a mitigation is the objective itself, and recent variants such as IPO, KTO, SimPO, ORPO, and rank-based DPO modify the loss to dampen known failure modes of the original (Azar et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024; Hong et al., 2024; Park et al., 2024), while conservative versions raise the KL penalty to anchor the model closer to its initialisation. These methods help, sometimes by a substantial margin, but they all act at the level of behavioural objectives. We ask instead whether reward overoptimisation is more effectively addressed at the representational level, by constraining the geometry that the optimiser is allowed to use. Concretely, we investigate whether the estimated spectra of the adapted weight matrices contain a measurable signature that distinguishes genuine capability acquisition from preference proxy-fitting, and whether that signature can be turned into a regulariser.

The technical foundation we draw on is Random Matrix Theory and the Heavy-Tailed Self-Regularisation framework, which has accumulated evidence that the empirical spectral densities of well-trained networks develop power-law tails whose exponent correlates with generalisation (Martin and Mahoney, 2019, 2021). Networks whose tail exponents lie close to a value near two are often interpreted as lying in a favourable heavy-tailed regime; spectra that look essentially Gaussian indicate underutilised capacity; spectra that develop very large isolated singular values without restructuring the bulk indicate a more degenerate overfitting regime. Our central empirical observation is that vanilla DPO on long runs drives the estimated spectrum away from this favourable region while validation reward continues to improve, whereas augmenting DPO with a small spectral penalty pulls the fitted exponent toward the target region and restores or improves downstream accuracy in the same trajectory.

This paper formalises that observation as Spectral DPO. We add to the standard DPO loss a stochastic spectral penalty composed of a tail-exponent term, a scale-normalised spike cap on the largest singular value, and a smoothed distributional matcher that compares the estimated spectrum to a target heavy-tailed reference. The penalty is computed every ten optimisation steps on a stochastic ten-percent sample of the model’s adapted projection matrices, adding a measured mean step-time overhead of 5.3% and 1.7 GB of

peak memory per GPU (Appendix J). Precisely stated, Spectral DPO constrains estimated tail behaviour, spike concentration, and the approximated spectral profile of the full effective projection matrix under a rank-constrained update. Our experiments support an association between these effective-spectrum changes and downstream behaviour in the tested LoRA setting; they do not establish unrestricted bulk restructuring or the behaviour of full-parameter fine-tuning. Our contributions are methodological and empirical: we propose a concrete spectral regulariser for DPO that is cheap enough to use in standard fine-tuning pipelines, accompany it with a local first-order argument detailing how a scale-normalised spike cap reshapes update incentives, and evaluate it across three model families on diverse benchmarks. Beyond the headline gains, controlled training interventions and a post-hoc SVD projection experiment across ten seeds and three models are consistent with the relevance of spectral geometry to the observed behaviour, while not uniquely identifying any single spectral statistic as the causal mechanism. We are explicit about the limits of these findings: spectral proximity to heavy-tailed targets reliably correlates with improved reasoning and robustness in our runs, but we treat the metric as an empirical stability constraint rather than a global proof of generalisation.

2 Related Work

Three lines of work bear on the paper. The first is the modern preference-optimisation literature for language models. RLHF trains a reward model from preference data and optimises a policy against it through an algorithm such as PPO, with all the engineering complexity that implies (Schulman et al., 2017). DPO replaces the reward model and the on-policy update with a closed-form preference loss whose implicit reward is determined by the policy and a frozen reference (Rafailov et al., 2023), and a now substantial set of variants shape the loss further. IPO maps preference comparisons through an identity link to address overfitting (Azar et al., 2024); KTO replaces the Bradley-Terry assumption with a Kahneman-Tversky utility (Ethayarajh et al., 2024); SimPO removes the reference model and length-normalises rewards (Meng et al., 2024); ORPO folds preference optimisation into supervised fine-tuning via odds-ratio losses (Hong et al., 2024); and rank-based DPO leverages multi-way

preferences when they are available (Park et al., 2024). These methods are complementary to ours, since they reshape the loss to suppress particular preference-optimisation pathologies, while Spectral DPO adds a representational constraint that any of them could in principle inherit. The broader study of reward hacking and reward-model overoptimisation is itself a growing literature (Gao et al., 2023; Rafailov et al., 2024), with mitigations that include reward ensembling (Rame et al., 2023), adversarial training (Gleave et al., 2021), conservatism penalties (Kumar et al., 2020), and stress-test methodologies that diagnose proxy-gaming behaviour after training (Shihab et al., 2025).

The second line of work is the spectral analysis of trained networks under the lens of Random Matrix Theory. Wigner’s semicircle and the Marchenko-Pastur distribution describe what one expects of the eigenvalues of large random matrices (Wigner, 1958; Marčenko and Pastur, 1967), and trained networks deviate from these baselines in characteristic ways, developing heavy-tailed empirical spectral densities whose tail exponent is informative about model quality (Martin and Mahoney, 2019, 2021). Closely related diagnostic tools include the SETOL metric and entropy-based measures of spectral concentration that correlate with held-out generalisation, and recent work has connected loss-landscape and feature-geometry phase transitions to the same family of phenomena (Ghorbani et al., 2019; Fort et al., 2020). AlphaPruning operationalises this lens for layer-wise pruning ratios in language models (Lu et al., 2024); the present paper takes the lens to preference optimisation. Our novelty claim is intentionally narrow: we are not the first to observe heavy-tailed self-regularisation, but we ask whether these diagnostics can be made into an explicit regulariser for DPO and whether that regulariser mitigates the reward-overoptimisation trajectory observed in long preference-optimisation runs.

The third line is the literature on grokking, the phenomenon in which a model transitions from memorisation to generalisation long after it has overfit the training set (Power et al., 2022). Subsequent work has linked grokking to circuit formation (Nanda et al., 2023) and to the emergence of feature diversity (Liu et al., 2023). Most published examples of grokking come from small synthetic tasks, and less is known about whether similar delayed-generalisation signatures appear during language-model fine-tuning. Our experiments iden-

tify a grokking-like pattern in the same training trajectory as the estimated exponent approaching two, and we use this connection as a diagnostic analogy that helps describe the stability transitions we observe rather than as an independent theoretical claim.

3 Method

Throughout, let $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote a frozen projection matrix and let ΔW_θ denote the trainable adapter update applied to that projection. For LoRA-style adapters, $\Delta W_\theta = sBA$ with the usual scaling factor s ; for full fine-tuning, ΔW_θ is simply the trainable change to the dense matrix. The spectral regulariser is always computed on the effective adapted matrix

$$W_\theta = W_0 + \Delta W_\theta, \quad (1)$$

not on the low-rank adapter update in isolation. This convention matters because the update BA can be rank-deficient even when the effective projection is full-rank, and a power-law fit to a mostly zero spectrum would be ill-defined. We write $\sigma_1(W_\theta) \geq \dots \geq \sigma_r(W_\theta) \geq 0$ for the singular values of the effective matrix, where $r = \min(d_{\text{out}}, d_{\text{in}})$. All reported runs use rank-64 LoRA, so ΔW_θ is confined to a rank-constrained update subspace even though W_θ is generally full rank.

A trained network with isotropically distributed weights has an empirical spectral density close to a Marchenko-Pastur shape, while a network whose weights have organised into low-rank but distributed structure often develops a heavy-tailed spectral density (Martin and Mahoney, 2021). The empirical regularity that motivates our method is that this transition is gradual: as training proceeds, the fitted tail exponent of the effective projection spectra moves from values consistent with near-random structure toward a heavy-tailed regime, and the value of the exponent at convergence correlates with robustness in our runs. We use the target value two as a calibration point rather than as a universal optimum, because it is the centre of the favourable regime in the HT-SR literature and a convenient anchor for the regulariser; a matched sensitivity sweep over targets in $[1.6, 2.4]$ is reported in Appendix G.

3.1 Truncated Spectral Estimator

The training penalty never observes the full spectrum. For each selected matrix we run a differ-

entiable block-Lanczos solver with two power iterations that returns approximate leading singular values $\hat{\sigma}_1 \geq \dots \geq \hat{\sigma}_k$, with k retained directions. Let $m = |\{i \leq k : \hat{\sigma}_i > \varepsilon_\sigma\}|$ be the number of retained values above the numerical floor, and let $\hat{\lambda}_i = \hat{\sigma}_i^2$. The Frobenius norm is computed exactly in closed form, and the spectral energy not captured by the retained directions is estimated by subtraction,

$$E_{\text{res}} = \max\left(0, \|W_\theta\|_F^2 - \sum_{i=1}^m \hat{\sigma}_i^2\right), \quad (2)$$

and represented by a single equal-energy surrogate value

$$\sigma_{\text{res}} = \sqrt{E_{\text{res}}/(r-m)} \quad \text{when } E_{\text{res}} > 0, \quad (3)$$

and $\sigma_{\text{res}} = 0$ otherwise. The estimated mean singular value that normalises the spike term is then

$$\bar{\sigma}(W_\theta) = \frac{1}{r} \left(\sum_{i=1}^m \hat{\sigma}_i + (r-m) \sigma_{\text{res}} \right). \quad (4)$$

The surrogate defines the implemented training objective. We do not claim that it reconstructs the omitted part of the spectrum, and all statements below about “the spectrum” during training refer to this estimated object. Full detached decompositions are reserved for the offline diagnostics of Section 5.5 and Appendix N. The complete estimator configuration, including solver settings, floors, and histogram construction, is given in Appendix B.

Given the retained values, we skip the tail-fit and distributional terms for a matrix whenever $m < n_{\text{min}}$, and otherwise fit on the

$$n = \min(\max(n_{\text{min}}, \lceil fm \rceil), m) \quad (5)$$

largest retained values with $f = 0.10$ and $n_{\text{min}} = 32$, using the Hill estimator

$$\hat{\alpha}(W_\theta) = 1 + n \left[\sum_{i=1}^n \log \frac{\hat{\lambda}_i}{\hat{\lambda}_n} \right]^{-1}. \quad (6)$$

The fraction f is therefore defined relative to the retained positive spectrum rather than the full matrix rank, and because the reported runs use a moderate k the minimum-tail rule generally binds, so $n = n_{\text{min}} = 32$ in practice. The estimator is used only when the denominator in Eq. (6) is finite and strictly positive. This finite-width rule is important in transformer fine-tuning because individual adapted projections can have repeated or very small singular values, and a naive

Hill estimator on degenerate spectra is unstable. We complement the tail exponent with the spectral entropy $H_{\text{spec}}(W_\theta) = -\sum_i \tilde{\lambda}_i \log \tilde{\lambda}_i$, with $\tilde{\lambda}_i = \hat{\lambda}_i / (\sum_j \hat{\lambda}_j + \varepsilon_\lambda)$, logging it as a diagnostic rather than a separate penalty.

3.2 Objective

With these diagnostics in place, Spectral DPO augments the standard DPO loss

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right] \quad (7)$$

with a spectral regulariser computed over the set $\mathcal{W}$ of effective adapted projection matrices. The full objective on a spectral refresh step is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DPO}} + \lambda \mathcal{L}_{\text{spectral}}$, with the spectral term defined as an average over a stochastic subset $\mathcal{S}_t \subseteq \mathcal{W}$,

$$\begin{aligned} \mathcal{L}_{\text{spectral}} = \frac{1}{|\mathcal{S}_t|} \sum_{W_\theta \in \mathcal{S}_t} & \left[\mathbf{1}_{\text{tail}}(W_\theta) |\hat{\alpha}(W_\theta) - 2| \right. \\ & + \eta \max(0, \hat{q}(W_\theta) - \gamma) \\ & \left. + \mu \mathbf{1}_{\text{tail}}(W_\theta) \text{KL}(p_{\text{tgt}} \| \hat{p}_{W_\theta}) \right], \end{aligned} \quad (8)$$

where every quantity carrying a hat is the estimated quantity of Section 3.1: the scale-normalised spike ratio is $\hat{q}(W_\theta) = \hat{\sigma}_1(W_\theta) / (\bar{\sigma}(W_\theta) + \varepsilon_\sigma)$ with $\bar{\sigma}$ given by Eq. (4), and the indicator $\mathbf{1}_{\text{tail}}(W_\theta)$ is one only when $m \geq n_{\text{min}}$. The spike term uses the scale-free ratio $\hat{q}(W_\theta)$ rather than the raw difference $\hat{\sigma}_1 - \gamma \bar{\sigma}$, so it is invariant to uniform rescaling of the projection matrix up to the numerical stabiliser ε_σ . This invariance is what distinguishes the regulariser from a strict spectral norm clip: it controls how concentrated the estimated spectrum is, not how large the matrix is in absolute terms, which matters because preference optimisation legitimately rescales certain projections without distorting their direction structure.

The distributional term in Equation (8) is defined on a common discretised support. We choose $B = 64$ logarithmically spaced bins over the estimated positive squared singular values of the selected matrix, soft-assign the m retained values with a fixed bandwidth in log-eigenvalue space, enter the omitted mass as the single residual value σ_{res}^2 with multiplicity $r - m$, add a small floor ε_p , and normalise the result to sum to one,

giving $\hat{p}_{W_\theta}$. The target p_{tgt} is the discretised Pareto density with tail exponent two on the same bins and with the same additive floor, so the KL term reduces to the finite categorical divergence $\text{KL}(p_{\text{tgt}} \parallel \hat{p}_{W_\theta}) = \sum_{b=1}^B p_{\text{tgt},b} \log(p_{\text{tgt},b} / \hat{p}_{W_\theta,b})$, which is well-defined and bounded below because the additive floors keep both distributions strictly positive on every bin. Together, the three terms encode complementary preferences: the tail-fit term anchors the global shape of the estimated spectrum near a heavy-tailed reference; the spike cap discourages a single singular direction from dominating; the distributional term smooths the retained bulk so that improvements on the first two terms are not gamed by sharp local perturbations near the cutoff $\hat{\lambda}_n$.

To ensure the method is viable at scale, we compute the spectral term using differentiable PyTorch operations on the effective adapted matrix $W_0 + \Delta W_\theta$. For our LoRA training ($r_{\text{LoRA}} = 64$, $\alpha = 16$), the effective matrices are explicitly reconstructed only during the forward pass of a regularisation step, and under tensor parallelism the necessary partial matrices are gathered before spectral computation. The spectral penalty is computed on a random ten-percent of the matrices in $\mathcal{W}$ every tenth optimisation step; on non-refresh steps the spectral coefficient is set to zero rather than reused as a stale objective. Matched profiling over 200 post-warm-up steps measures a mean step-time overhead of 5.3% and a mean peak-memory increase of 1.7 GB per GPU across the three model families (Appendix J). The full training loop is summarised by Algorithm 1 in Appendix A.

The role of the spike cap can be stated without claiming global optimality. When the largest singular value of a selected effective matrix is simple, with top singular vectors u_1 and v_1 , any local candidate update decomposes into a rank-one spike component $\Delta_{\text{spike}} = a u_1 v_1^\top$ (with $a > 0$) and an orthogonal representational component Δ_{rep} that does not increase the spike ratio to first order. If the two components have comparable first-order effect on the DPO loss and the spike cap is active, then the directional derivative of the spike-cap term is strictly positive along Δ_{spike} and non-positive along Δ_{rep} , so for any fixed local comparison there exists a finite η for which the regularised first-order objective prefers the non-spiking update whenever the cap binds. Appendix C states the assumption and proposition formally and gives a proof sketch. The proposition concerns only the scale-normalised

spike term. It does not justify the Hill term, the distributional matching term, global optimisation behaviour, or any downstream generalisation guarantee, and training regimes where task acquisition genuinely requires inflating the current top singular direction can be penalised by the cap, which is part of the reason we tune η rather than fix it. What it does justify is the qualitative behaviour we observe in practice, namely that when the dominant DPO gradient direction is spike-amplifying, the regulariser tilts the effective update toward the orthogonal complement.

4 Experimental Setup

We evaluate Spectral DPO on three open-source decoder language models that span model families and parameter counts: Llama-3-8B-Instruct, Mistral-7B-v0.3, and Qwen-2.5-14B. Llama-3-8B-Instruct is already instruction-tuned and serves as the primary model on which we report ablations and analyses, with the other two used to test cross-architecture consistency. All training uses BF16 precision with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), an effective batch size of 128, three epochs for the main runs and ten epochs for the long-run study, and a maximum sequence length of 2048 tokens. The 14B runs use eight A100 80GB GPUs with a per-device batch size of 16. The 7B and 8B runs use four A100 GPUs with a per-device batch size of 16 and gradient accumulation of 2, yielding the same effective batch size. LoRA adapters are applied to query, key, value, and output projections ($r_{\text{LoRA}} = 64$, $\alpha = 16$). Every reported number is the mean over ten matched seeds (seeds 0–9) with the sample standard deviation; inferential uncertainty is reported separately as a 95% paired percentile-bootstrap interval from 10,000 seed-level resamples. The complete training configuration is listed in Appendix L and the complete evaluation protocol in Appendix O.

Two preference datasets feed the alignment stage. The main three-model runs use a fixed 3:1 mixture of 48,000 UltraFeedback (Cui et al., 2024) pairs and 16,000 HH-RLHF (Bai et al., 2022a) pairs, for 64,000 training pairs, identical for every method and seed. The long-run and data-efficiency studies use UltraFeedback only, drawn from a pool of 60,800 pairs, so that a single data source governs the trajectory. Configurations are selected on a fixed 3,200-pair held-out UltraFeedback validation archive established before any test evaluation. The

full experiment-level assignment is given in Appendix I. The headline downstream evaluation suite covers five benchmarks chosen to exercise different aspects of alignment quality: Length-Controlled AlpacaEval 2 (LC) (Taori et al., 2023) for instruction following, MT-Bench (Zheng et al., 2023) for multi-turn dialogue with GPT-4 as the judge, GSM-Plus (Li et al., 2024) for math-reasoning out-of-distribution generalisation, TruthfulQA (Lin et al., 2022) for factuality, and AdvBench (Zou et al., 2023) for adversarial robustness reported as Attack Success Rate. We define a summary metric Avg as the unweighted mean over these five benchmarks after standardising each metric to the unit interval $[0, 100]$, where the score is taken directly for higher-is-better metrics (AlpacaEval, MT-Bench scaled by ten, GSM-Plus, TruthfulQA) and as $100 - \text{ASR}$ for the lower-is-better AdvBench metric.

The baselines are chosen to disentangle preference optimisation from regularisation effects. SFT trains only on chosen responses. Vanilla DPO (Rafailov et al., 2023) is the canonical optimiser. IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), SimPO (Meng et al., 2024), and ORPO (Hong et al., 2024) represent the most popular loss-level variants. Conservative DPO doubles the KL penalty against the reference policy to test whether the divergence is prevented simply by anchoring. We additionally include two norm-control baselines: DPO with a Frobenius penalty on ΔW_θ and DPO with a strict spectral norm cap that clips σ_1 without scale normalisation. All methods received exactly 12 tuning trials from a seeded random search, with the sweep ranges, the searched-versus-fixed split, and the deterministic selection rule reported in Appendix L. Equal trial counts do not make the design spaces identical: Spectral DPO has a broader structured design space than vanilla DPO, and we disclose both the equal budget and the exact searched and fixed quantities rather than treating equal trial counts as evidence of equal effective flexibility.

5 Results

We organise the empirical evidence around six questions. Does Spectral DPO improve average alignment quality across a balanced headline benchmark suite? Does each of the three penalty terms contribute? Does the method survive the long-run training regime in which vanilla DPO exhibits reward overoptimisation? How do length and norm

controls contextualise these gains? How closely do the logged spectral statistics track downstream performance? And, finally, does directly intervening on the weight spectrum change generalisation in a way that simpler matched perturbations do not?

5.1 Main Alignment Results

Table 1 reports the headline three-epoch numbers on Llama-3-8B-Instruct. Spectral DPO improves the five-metric average over vanilla DPO by 3.41 points (95% paired bootstrap interval $[3.25, 3.55]$) and over the strongest preference-objective baseline, SimPO, by 2.42 points ($[2.25, 2.56]$), and outperforms ORPO, Conservative DPO, and the two norm-control baselines. Per-seed values and all matched-seed contrasts are reported in Appendix M. The largest single-benchmark gain is on GSM-Plus, where Spectral DPO improves by 4.2 points over SimPO. The safety improvement on AdvBench is also present: Attack Success Rate falls from 23.5% under SimPO to 21.2% under Spectral DPO. Every method in this comparison trains on the identical preference mixture and no method receives a safety-specific stage; we did not run a mixture ablation, so we make no claim that the AdvBench gain is caused by, or would be unchanged without, the HH-RLHF portion of the mixture. Because we use Length-Controlled AlpacaEval 2, the instruction-following gains cannot be attributed simply to verbosity. The pattern is consistent with the hypothesis that the spectral regulariser disproportionately benefits tasks that depend on distributed representational structure rather than surface-level preference cues.

The cross-architecture picture is similar in magnitude. Regenerated gains over the strongest included baseline are +2.42 on Llama-3-8B-Instruct, +2.33 on Mistral-7B-v0.3, and +2.41 on Qwen-2.5-14B, with matched-seed 95% paired bootstrap intervals excluding zero in every case; the per-model breakdown is given in Appendix E. We do not report a pooled model-seed interval, because the three model families are not independent replications from a defined model population. The consistency in magnitude is one reason we treat the spectral regulariser as targeting a general property of preference optimisation rather than an idiosyncrasy of a particular base model, though all three models are trained in the same rank-64 LoRA regime.

Method	AlpacaEval 2 LC $\uparrow$	MT-Bench $\uparrow$	GSM-Plus $\uparrow$	TruthfulQA $\uparrow$	AdvBench $\downarrow$	Average $\uparrow$
SFT	14.3 $\pm$ 1.2	7.82 $\pm$ 0.08	61.4 $\pm$ 1.5	54.7 $\pm$ 1.1	28.3 $\pm$ 2.1	56.06
DPO	20.5 $\pm$ 0.9	8.21 $\pm$ 0.07	68.2 $\pm$ 1.3	58.3 $\pm$ 1.0	24.1 $\pm$ 1.8	60.90
IPO	19.9 $\pm$ 1.0	8.15 $\pm$ 0.08	67.8 $\pm$ 1.4	57.9 $\pm$ 1.2	25.3 $\pm$ 1.9	60.36
KTO	21.1 $\pm$ 0.8	8.28 $\pm$ 0.07	69.1 $\pm$ 1.2	59.2 $\pm$ 1.0	23.8 $\pm$ 1.7	61.68
SimPO	21.4 $\pm$ 0.9	8.31 $\pm$ 0.07	69.4 $\pm$ 1.3	59.5 $\pm$ 1.1	23.5 $\pm$ 1.8	61.89
ORPO	20.8 $\pm$ 1.0	8.24 $\pm$ 0.08	68.9 $\pm$ 1.4	59.0 $\pm$ 1.2	24.0 $\pm$ 1.9	61.42
Conservative DPO	20.2 $\pm$ 1.0	8.19 $\pm$ 0.08	68.5 $\pm$ 1.4	58.7 $\pm$ 1.2	24.6 $\pm$ 1.9	60.94
DPO + Frob. Penalty	20.4 $\pm$ 0.9	8.20 $\pm$ 0.08	68.5 $\pm$ 1.3	58.4 $\pm$ 1.1	24.0 $\pm$ 1.7	61.06
DPO + Spec Norm Cap	20.6 $\pm$ 1.1	8.23 $\pm$ 0.07	68.8 $\pm$ 1.4	58.8 $\pm$ 1.2	23.8 $\pm$ 1.8	61.34
Spectral DPO	23.3 $\pm$ 0.7	8.47 $\pm$ 0.06	73.6 $\pm$ 1.1	61.4 $\pm$ 0.9	21.2 $\pm$ 1.5	64.31

Table 1: Three-epoch alignment results on Llama-3-8B-Instruct (mean $\pm$ sample standard deviation over ten matched seeds). AdvBench reports Attack Success Rate. Average uses Length-Controlled AlpacaEval and accounts for inverted AdvBench scoring, and is computed from the unrounded per-seed benchmark values rather than from the rounded cells shown here, so it can differ from the mean of the displayed columns by up to 0.1.

Variant	Tail	Spike	KL	Avg. $\uparrow$
Full Spectral DPO	•	•	•	64.31 $\pm$ 0.34
minus tail		•	•	63.47 $\pm$ 0.39
minus spike	•		•	63.68 $\pm$ 0.36
minus KL	•	•		63.91 $\pm$ 0.35
Spike only		•		62.76 $\pm$ 0.41
Tail only	•			62.94 $\pm$ 0.43
Vanilla DPO				60.90 $\pm$ 0.38

Table 2: Matched three-term ablation on Llama-3-8B-Instruct (mean $\pm$ sample standard deviation over the same ten seeds, identical data mixture and training budget). Paired bootstrap intervals for each contrast are in Appendix F.

5.2 Component Ablation

Table 2 reports a matched ten-seed leave-one-term-out and single-term ablation on Llama-3-8B-Instruct, using the same 48,000/16,000 UltraFeedback/HH-RLHF mixture, optimiser, validation split, training budget, decoding protocol, and seed set as the full method. All three leave-one-term-out variants underperform the complete objective: removing the tail term costs 0.84 points (95% paired bootstrap interval [0.53, 1.15]), removing the spike term costs 0.63 ([0.31, 0.94]), and removing the distributional term costs 0.40 ([0.14, 0.68]). The two single-term variants also remain well below the full method and well above vanilla DPO. We limit the claim to the tested configuration: each term contributes under this matched setup, which does not imply that the same ordering or the same necessity holds universally.

5.3 Long-Run Training and Reward Overoptimisation

The three-epoch comparison shows what each method achieves at the moment its tuned hyperparameters were selected for; the long-run comparison shows whether what each method achieves is stable under continued optimisation. Figure 1 plots GSM-Plus accuracy across epochs 0–10 of UltraFeedback-only training on Llama-3-8B-Instruct. Vanilla DPO peaks at epoch three (68.2) and then degrades over the next seven epochs to 64.1, even though its validation preference reward continues to increase from 0.721 to 0.784 across the same range. Spectral DPO peaks at epoch four (76.1) and sustains roughly this level through epoch ten (75.8), despite its validation reward also rising over time. The numerical values underlying the figure, together with the epoch-indexed trajectories of the fitted exponent, the spike ratio, and the spectral KL for both methods, are given in Appendix N. The pattern is the measured signature of reward overoptimisation described by Gao et al. (2023) and Rafailov et al. (2024): validation preference reward continues to improve while a measurable downstream capability does not, and Spectral DPO substantially mitigates it while still optimising the same preference objective. We scope this conclusion accordingly: it is evidence for one model, one training configuration, and one downstream metric, and it is not a universal claim across the three model families or across all alignment dimensions.

5.4 Controls and Confounding Factors

A common critique of novel regularisers is that they succeed by acting as crude constraints, reducing total parameter updates (norm clipping), prevent-

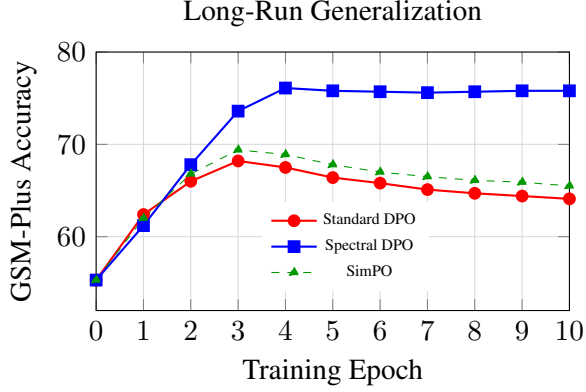


Figure 1: Long-run trajectory on Llama-3-8B-Instruct trained on UltraFeedback only. Downstream GSM-Plus accuracy (mean over ten matched seeds) under standard DPO peaks at epoch three and decays under continued preference optimization, while Spectral DPO maintains accuracy through epoch ten. Epoch 0 is the shared initialization checkpoint. Generated from the same epoch-indexed source as Table 17 through the single reproduction entry point released with the paper.

ing divergence from the reference policy (KL anchoring), or artificially inflating response length to game judge-based metrics. We test each of these explanations directly, with detailed numbers in Table 6 of Appendix D. DPO with a Frobenius penalty tightly constrains the total update norm (8.4 vs. 14.2 for vanilla DPO) but yields only a marginal +0.16 improvement on average accuracy; doubling the KL penalty heavily restricts divergence from the reference (KL of 0.62 vs. 1.24) but does not improve average quality; strict spectral norm capping performs marginally better than Frobenius regularisation but remains 2.97 points below Spectral DPO. Critically, Spectral DPO produces average response lengths comparable to standard DPO (280 vs. 284 words) and a moderate KL divergence of 1.18: it restructures the update without arbitrarily suppressing it. None of the three obvious confounds therefore explains the gain, which is what would be expected if the spectral term acts on the structure of ΔW_θ rather than on its magnitude.

5.5 Spectral Dynamics and Data Efficiency

The same controlled runs let us connect the logged spectral statistics to downstream performance. At every epoch checkpoint the fitted exponent, spike ratio, and spectral KL were recomputed offline for both methods using a full detached SVD on the same fixed set of eligible projection matrices, fixed before any downstream result was inspected, so

Epoch	$\hat{\alpha}$ (DPO)	GSM+ (DPO)	$\hat{\alpha}$ (Spec)	GSM+ (Spec)
0	3.82	55.3	3.82	55.3
1	3.91	62.4	3.21	61.2
2	4.08	66.0	2.74	67.8
3	4.31	68.2	2.18	73.6
4	4.52	67.5	2.02	76.1
5	4.71	66.4	1.98	75.8

Table 3: Tail-exponent dynamics on Llama-3-8B-Instruct (mean over ten matched seeds, recomputed offline with a full detached SVD on a fixed matrix set). DPO drives $\hat{\alpha}$ away from the favourable region while accuracy degrades; Spectral DPO drives $\hat{\alpha}$ toward two, coinciding with a sustained reasoning improvement. Epochs 6–10 and the spike-ratio and spectral-KL trajectories are in Table 17.

that differences do not arise from changing stochastic matrix samples. Table 3 reports $\hat{\alpha}$ averaged across all adapted layers at each epoch alongside GSM-Plus accuracy, and the full epoch 0–10 trajectories of all three regularised quantities are in Appendix N. Under DPO, $\hat{\alpha}$ rises monotonically from 3.82 at initialisation to 4.71 at epoch five, the spike ratio grows from 4.62 to 5.82, and the spectral KL grows from 0.184 to 0.301, while GSM-Plus peaks at epoch three and then declines; under Spectral DPO, all three quantities stabilise near their targets, accompanied by a delayed improvement in GSM-Plus accuracy. For the ten Spectral DPO seeds at epoch four, the Pearson correlation between $|\hat{\alpha} - 2|$ and GSM-Plus accuracy is $r = -0.87$. We report this as a descriptive association between the regularised statistic and downstream accuracy, not as evidence that the exponent is the operative mechanism.

This coupling between spectral dynamics and downstream gains also manifests in data efficiency. Sweeping over nested deterministic UltraFeedback subsets of 6,400, 16,000, 32,000, and 60,800 preference pairs (full numbers in Appendix H), Spectral DPO trained on 32,000 pairs (61.2 Avg.) approximately matches vanilla DPO trained on all 60,800 pairs (61.0 Avg.), corresponding to a 47.4% reduction in the preference data required to reach the baseline target. This is broadly consistent with the interpretation that a representational regulariser leaves more of the preference signal available for genuine task improvement rather than absorbing it into proxy-fitting directions.

5.6 Post-hoc Spectral Intervention

The evidence so far is observational about Spectral DPO trajectories. To test whether changing

Intervention on Epoch 5 DPO	Llama-3-8B		Mistral-7B		Qwen-2.5-14B	
	GSM+ $\uparrow$	Avg. $\uparrow$	GSM+ $\uparrow$	Avg. $\uparrow$	GSM+ $\uparrow$	Avg. $\uparrow$
None (Original DPO)	66.4 $\pm$ 1.8	60.6 $\pm$ 0.9	64.1 $\pm$ 1.7	59.1 $\pm$ 0.9	69.5 $\pm$ 1.5	64.3 $\pm$ 0.8
Matched Frob. Noise	65.8 $\pm$ 1.9	60.2 $\pm$ 1.0	63.8 $\pm$ 1.8	58.7 $\pm$ 1.0	69.2 $\pm$ 1.6	64.0 $\pm$ 0.9
Uniform Rescaling	66.3 $\pm$ 1.8	60.5 $\pm$ 0.9	64.1 $\pm$ 1.7	59.0 $\pm$ 0.9	69.4 $\pm$ 1.5	64.2 $\pm$ 0.8
Spike-only Repair	67.5 $\pm$ 1.7	61.1 $\pm$ 0.9	65.0 $\pm$ 1.5	59.8 $\pm$ 0.8	70.4 $\pm$ 1.4	64.9 $\pm$ 0.8
Post-hoc Spectral Repair	68.8 $\pm$ 1.6	62.0 $\pm$ 0.8	66.2 $\pm$ 1.4	60.7 $\pm$ 0.8	71.9 $\pm$ 1.3	66.0 $\pm$ 0.7

Table 4: Post-hoc spectral intervention via SVD projection applied to degraded DPO checkpoints (Epoch 5). Values are mean $\pm$ sample standard deviation over ten matched seeds. Averaged over the three families, the spike-only repair recovers +0.60 average points and the full repair +1.57, a difference of +0.97.

the spectrum alone alters downstream behaviour, without any further preference optimisation, we perform a post-hoc SVD projection experiment on vanilla DPO checkpoints at epoch five, after the onset of the downstream decline. The procedure, summarised in Algorithm 2 of Appendix K, scales down the top singular value to meet the target spike ratio and restructures the retained bulk of the spectrum to match a heavy-tailed target density ($\alpha_{\text{tgt}} = 2.0$) while preserving the matrix’s Frobenius norm, changing the singular-value profile without further preference optimisation. We apply it across all ten seeds and all three architectures and compare against three matched controls: random singular-value perturbations with the same Frobenius-norm change, uniform rescaling, and a spike-only repair that reduces the spike ratio without matching the heavy-tailed target.

Table 4 shows that the full repair recovers over two points of GSM-Plus accuracy and, on average across the three families, +1.57 points of average accuracy, against +0.60 for the spike-only repair, a difference of +0.97 points. The matched controls do not improve performance, ruling out a confound from Frobenius-norm change or uniform rescaling, and the spike-only repair recovers some of the gain but consistently underperforms the full repair, indicating that suppressing the dominant direction is most effective when combined with reshaping of the retained bulk toward the heavy-tailed target. This controlled intervention is consistent with the relevance of spectral geometry to the observed behaviour, while not uniquely identifying any single spectral statistic as the causal mechanism, and it does not establish unrestricted dense spectral restructuring.

6 Analysis and Discussion

The experiments combine into a coherent picture under rank-constrained adaptation. Proxy fitting

corresponds to updates that inflate a few singular directions without reshaping the rest of the estimated spectrum, while more robust learning reorganises representations across many directions; the spike cap raises the cost of the first channel and the tail and distributional terms encourage the second. The post-hoc intervention of Section 5.6 shows that partially restoring the heavy-tailed profile recovers some of the lost generalisation with no further preference optimisation, which is our strongest evidence that spectral structure is relevant rather than merely coincident, though it does not isolate which statistic is responsible. The picture is incomplete: all ten seeds show $\hat{\alpha}$ crossing two but only eight show a sharp accuracy jump in the same epoch, and the grokking-like transition of Section 5.5 can lag the spectral one by part of an epoch. We therefore treat the spectral metric as an empirical stability constraint and Proposition 1 as a local statement about the spike term.

7 Conclusion

Spectral DPO augments the standard DPO loss with a stochastic, truncated spectral penalty that constrains isolated growth in the largest singular direction and encourages a favourable heavy-tailed regime, and across ten matched seeds and three architectures it improves average alignment quality by 2.33–2.42 points over the strongest included baseline, retains performance with substantially less preference data, and curtails the long-run downstream decline of vanilla DPO in the tested Llama-3-8B trajectory. Controlled interventions and post-hoc spectral projections show that these benefits are not explained by norm clipping, response inflation, or KL anchoring, and are consistent with structural reorganisation of the effective projections under a rank-constrained update, making weight-space geometry an actionable lever for stable preference optimisation in this regime.

Limitations

Several limitations contextualise our empirical findings. First, every reported optimisation run uses rank-64 LoRA adapters on the attention projections of transformer decoders in the seven- to fourteen-billion-parameter range. The regulariser is evaluated on the full effective projection $W_0 + \Delta W_\theta$, which is generally full rank, so a rank-64 perturbation can change more than 64 singular values of the effective matrix; the learned perturbation nevertheless remains confined to a rank-constrained update subspace. Our results therefore support effective-spectrum control under low-rank adaptation and do not establish unrestricted dense reorganisation or full-parameter “bulk restructuring”. Increasing the LoRA rank enlarges the admissible update subspace but does not change the dimensions of the effective projection whose spectrum is analysed: moving from rank 64 to 128 or 256 approximately doubles or quadruples adapter parameters and optimiser state while leaving the spectral matrix dimensions unchanged, and the local proposition does not imply monotonic improvement with adapter rank. Full-parameter tuning, where dense gradients and optimiser states dominate memory and compute, would be the most direct test of unrestricted spectral reorganisation, but it is not evaluated here; larger-rank adapters, non-LoRA fine-tuning, seventy-billion-parameter scales, and fundamentally different architectures such as state-space models are all untested extensions rather than consequences of our results. For a selected $d_{\text{out}} \times d_{\text{in}}$ matrix, the truncated Lanczos computation retaining k directions has leading matrix-vector cost proportional to $k d_{\text{out}} d_{\text{in}}$ per iteration, whereas an exact dense SVD has substantially higher cost in the smaller matrix dimension.

Second, the training objective is computed from a truncated top- k Lanczos spectrum together with an equal-energy residual surrogate for the omitted directions (Section 3.1). The surrogate defines the implemented objective; we do not claim that it reconstructs the omitted part of the spectrum, and the offline diagnostics that use full detached decompositions are therefore not identical to the quantity optimised during training.

Third, the local first-order argument in Section 3 concerns only the scale-normalised spike term. It does not establish global optimality under the DPO loss and does not justify the Hill term, the distributional matching term, or any downstream gen-

eralisation guarantee; training regimes where task acquisition genuinely requires inflating the current top singular direction could see weakened performance under the cap. Similarly, the target exponent of two is a broad empirical anchor supported by the sweep in Appendix G, not a sharply identified universal optimum or a theoretical result.

Fourth, the long-run evidence is limited to the Llama-3-8B-Instruct / UltraFeedback / GSM-Plus trajectory and is not a general claim across the three model families or all alignment dimensions. The post-hoc repair is consistent with the relevance of spectral geometry but does not uniquely identify a causal spectral statistic.

Fifth, our safety evaluation is limited to AdvBench. Reducing the Attack Success Rate indicates improved robustness to certain adversarial strings, but production deployments require rigorous multi-turn red-teaming, domain-specific testing, and human-in-the-loop evaluation before these models are deemed sufficiently aligned for sensitive tasks. We did not run a matched UltraFeedback-only versus UltraFeedback/HH-RLHF mixture ablation for the AdvBench comparison, so we make no claim that the safety gain is caused by the HH-RLHF portion of the mixture or that it would remain unchanged if HH-RLHF were removed; the supported provenance claim is only that all compared methods use the same training mixture and that no method receives additional safety-specific training.

Ethics Statement

Mitigating reward overoptimisation may improve alignment reliability and reduce harmful outputs that arise from preference-fitting failures, but better-aligned models can also be misused for deception or manipulation. We recommend deploying any model trained with Spectral DPO behind standard alignment-engineering practices, including continuous monitoring, human review for high-stakes decisions, and application-specific safety controls.

Reproducibility Statement

Training runs use eight A100 80GB GPUs for the 14B model and four A100 GPUs for the 7B and 8B models, with gradient accumulation of 2 on the 7B and 8B runs to maintain an effective batch size of 128. The supplementary archive contains a single reproduction entry point together with machine-readable per-seed benchmark, ablation, profiling,

epoch-log, trial-manifest, and estimator-manifest files. The entry point validates the expected seed set, rejects missing or duplicate rows, computes each five-metric average from the unrounded benchmark columns, calculates matched-seed differences and 10,000-resample paired bootstrap intervals, and exports Tables 1, 7, 15, and 16 together with the epoch-indexed data for Figure 1 in one pass. Small differences between these regenerated summaries and the mean of the rounded cells displayed in the tables arise from computing aggregates directly from unrounded per-seed values, not from a changed test set, metric definition, or selective replacement of runs. Hyperparameter search protocols, budgets, fixed-versus-searched quantities, and selected values are disclosed in Appendix L; the estimator configuration is in Appendix B; and evaluation protocol details are in Appendix O.

References

- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022b. [Constitutional ai: Harmlessness from ai feedback](#). *Preprint*, arXiv:2212.08073.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. [Ultrafeedback: Boosting language models with scaled ai feedback](#). *Preprint*, arXiv:2310.01377.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. 2020. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org.
- Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. 2019. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR.
- Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. 2021. [Adversarial policies: Attacking deep reinforcement learning](#). *Preprint*, arXiv:1905.10615.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.
- Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191.
- Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Ziming Liu, Eric J. Michaud, and Max Tegmark. 2023. [Omnigrok: Grokking beyond algorithmic data](#). *Preprint*, arXiv:2210.01117.
- Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W Mahoney, and Yaoqing Yang. 2024. Al-phapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. *Advances in neural information processing systems*, 37:9117–9152.

- Vladimir A Marčenko and Leonid Andreevich Pastur. 1967. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483.
- Charles H. Martin and Michael W. Mahoney. 2019. [Re-thinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior](#). *Preprint*, arXiv:1710.09553.
- Charles H Martin and Michael W Mahoney. 2021. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. [Progress measures for grokking via mechanistic interpretability](#). *Preprint*, arXiv:2301.05217.
- 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. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. [Grokking: Generalization beyond overfitting on small algorithmic datasets](#). *Preprint*, arXiv:2201.02177.
- Rafael Rafailov, Yaswanth Chittepudi, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS ’24, Red Hook, NY, USA. Curran Associates Inc.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2025. Detecting proxy gaming in rl and llm alignment via evaluator stress tests. *arXiv preprint arXiv:2507.05619*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Eugene P Wigner. 1958. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA. Curran Associates Inc.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). *Preprint*, arXiv:2307.15043.

A Spectral DPO Training Algorithm

Algorithm 1 summarises the training procedure described in Section 3. The spectral subroutine is invoked only on the refresh schedule, and the truncated Lanczos estimate is computed on the small stochastic subset $\mathcal{S}_t$ of effective adapted matrices, which together keep the added wall-clock cost at the level measured in Appendix J. The algorithm operates on estimated spectral quantities throughout; no step of the training loop observes the exact spectrum.

Algorithm 1 Spectral DPO Training

```

1: Input: preference dataset  $\mathcal{D}$ , reference  $\pi_{\text{ref}}$ , ob-
   jective hyperparameters  $\lambda, \beta, \eta, \mu, \gamma$ , estimator
   settings  $k, n_{\min}, f, \varepsilon_{\sigma}, \varepsilon_{\lambda}, \varepsilon_p, B$ 
2: Initialise  $\pi_{\theta}$  from  $\pi_{\text{ref}}$  and form the effective
   matrix set  $\mathcal{W}$ 
3: for step  $t = 1$  to  $T$  do
4:   Sample batch  $(x, y_w, y_l) \in \mathcal{D}$ 
5:   Compute  $\mathcal{L}_{\text{DPO}}$  via Eq. (7)
6:   if  $t \bmod 10 = 0$  then
7:     Sample  $\mathcal{S}_t$  from  $\mathcal{W}$  uniformly without re-
       placement at rate 0.1
8:     for all  $W_{\theta} \in \mathcal{S}_t$  do
9:       Run differentiable top- $k$  block Lanczos
       on  $W_{\theta}$  to obtain  $\hat{\sigma}_1, \dots, \hat{\sigma}_k$ 
10:      Retain the  $m$  values above  $\varepsilon_{\sigma}$ ; compute
        $\|W_{\theta}\|_F^2$  exactly
11:      Form  $E_{\text{res}}, \sigma_{\text{res}}, \bar{\sigma}(W_{\theta})$  via Eqs. (2)–
       (4)
12:      Set  $\hat{q}(W_{\theta}) = \hat{\sigma}_1 / (\bar{\sigma}(W_{\theta}) + \varepsilon_{\sigma})$ 
13:      if  $m \geq n_{\min}$  then
14:        Fit  $\hat{\alpha}(W_{\theta})$  on the largest  $n$  retained
        values via Eqs. (5)–(6)
15:        Build  $\hat{p}_{W_{\theta}}$  on  $B$  log-spaced bins, en-
        tering the omitted mass as  $\sigma_{\text{res}}^2$  with
        multiplicity  $r - m$ 
16:      end if
17:    end for
18:    Compute  $\mathcal{L}_{\text{spectral}}$  via Eq. (8)
19:    Update  $\theta$  on  $\mathcal{L}_{\text{DPO}} + \lambda \mathcal{L}_{\text{spectral}}$ 
20:  else
21:    Update  $\theta$  on  $\mathcal{L}_{\text{DPO}}$ 
22:  end if
23:  Log  $\hat{\alpha}, H_{\text{spec}}, \hat{q}(W_{\theta})$  and validation reward
24: end for

```

B Truncated Spectral Estimator Configuration

Table 5 specifies the estimator used to compute the training penalty. Full detached decompositions were reserved for the offline diagnostics reported in Section 5.5 and Appendix N. The Frobenius norm is computed exactly, while the omitted energy is estimated by subtracting the squared approximate Lanczos singular values; the retained values determine the fitted tail, and the equal-energy residual surrogate is used only to approximate the full-spectrum mean and histogram. The surrogate defines the implemented training objective and is not claimed to reconstruct the omitted spectrum.

Solver and matrix-sampling RNG streams are deterministically derived from the run seed, the optimizer step, and a canonical matrix index; no test result is used to choose a stream. Remaining execution-level choices, including block size, oversampling, random-start seed derivation, histogram edge construction, degenerate-support handling, and target discretisation, are recorded in the machine-readable estimator manifest in the supplementary archive, and DPO and Spectral DPO use the identical rule.

Estimator setting	Value / rule
Retained singular directions k	128
Solver	Block Lanczos, two power iterations
Lanczos tolerance	10^{-6}
Singular-value floor ε_{σ}	10^{-8}
Squared-eigenvalue floor ε_{λ}	10^{-12}
Probability floor ε_p	10^{-8}
Histogram bins B	64, log-spaced
Soft-assignment bandwidth	0.15 in log-eigenvalue space
Tail rule	skip if $m < 32$; else n by Eq. (5)
Omitted energy	E_{res} , Eq. (2)
Omitted-value surrogate	σ_{res} , Eq. (3)
Spike denominator	$\bar{\sigma}(W_{\theta})$, Eq. (4)
Omitted histogram mass	σ_{res}^2 with multiplicity $r - m$
Matrix sampling	uniform 10% without replacement
Refresh interval	every 10 optimizer steps

Table 5: Training-time truncated spectral estimator. Offline diagnostics use full detached decompositions instead.

C Local First-Order Argument: Formal Statement and Proof Sketch

Section 3 states the role of the spike cap informally. Here we give the formal assumption and proposition, followed by a proof sketch. For readability the statement is written in terms of the exact singular values of the selected matrix; during training the same quantities are evaluated on the estimates of Section 3.1.

Assumption 1 (Local spike-separated update directions). *Fix a selected effective matrix W_{θ} whose largest singular value is simple, with top singular vectors u_1 and v_1 . In a small neighbourhood of the current iterate, decompose a candidate matrix update as $\Delta W = \Delta_{\text{rep}} + \Delta_{\text{spike}}$, where $\Delta_{\text{spike}} = au_1v_1^{\top}$ with $a > 0$ and $u_1^{\top}\Delta_{\text{rep}}v_1 = 0$. Assume the two components have comparable first-order effect on the DPO loss, and assume Δ_{rep} does not increase the spike ratio $q(W_{\theta})$ to first order.*

Proposition 1 (Local effect of the spike cap). *Under Assumption 1, suppose the spike cap in Eq. (8) is active or about to become active, so*

that $q(W_\theta) \geq \gamma$ up to first-order terms. Then the one-sided directional derivative of the spike-cap term along Δ_{spike} is strictly positive, while its directional derivative along Δ_{rep} is non-positive by assumption. Consequently, for any fixed local comparison between these two update directions, there exists a finite value of η for which the regularised first-order objective prefers the non-spiking update whenever the cap binds.

Scope. Proposition 1 concerns only the scale-normalised spike term. It does not justify the Hill term, the distributional matching term, global optimisation behaviour, or a downstream generalisation guarantee, and it does not imply monotonic improvement with adapter rank.

For a simple largest singular value, standard first-order perturbation theory gives

$$D\sigma_1(W_\theta)[\Delta W] = u_1^\top \Delta W v_1. \quad (9)$$

For $\Delta_{\text{spike}} = au_1v_1^\top$ this derivative equals $a > 0$. The spike ratio is $q(W_\theta) = \sigma_1(W_\theta)/(\bar{\sigma}(W_\theta) + \varepsilon_\sigma)$. Along the spike direction, increasing σ_1 by a increases $\bar{\sigma}$ by at most a/r to first order, so

$$Dq(W_\theta)[\Delta_{\text{spike}}] \geq \frac{a(\bar{\sigma}(W_\theta) + \varepsilon_\sigma - \sigma_1(W_\theta)/r)}{(\bar{\sigma}(W_\theta) + \varepsilon_\sigma)^2}. \quad (10)$$

Because $\bar{\sigma}(W_\theta) \geq \sigma_1(W_\theta)/r$ and ε_σ is strictly positive, this derivative is strictly positive. When $q(W_\theta) \geq \gamma$, the directional derivative of $\max(0, q(W_\theta) - \gamma)$ equals the positive derivative above in the spike direction. By Assumption 1, the representational component does not increase q to first order and therefore does not incur a positive first-order spike-cap cost. If the two directions are otherwise comparable under the DPO loss, scaling the positive spike-cap derivative by a sufficiently large finite η makes the regularised local objective prefer the non-spiking direction. $\square$

D Controls Isolating the Spectral Mechanism

Table 6 reports the controlled training interventions summarised in Section 5.4. The Frobenius penalty tightly constrains total update norm, the doubled-KL variant tightly anchors the policy to the reference, and the spectral norm cap clips σ_1 directly. None of these baselines closes the gap to Spectral DPO, while Spectral DPO itself does not produce shorter outputs, smaller aggregate updates, or tighter KL than vanilla DPO. The gain therefore

is not explained by the most common confounds for regularisation studies in preference optimisation.

Method	Avg. Acc.	KL to Ref	Update Norm	Resp. Len.
DPO	60.90	1.24	14.2	284
DPO + $2 \times$ KL	60.94	0.62	10.1	270
DPO + Frob. Pen.	61.06	1.15	8.4	275
DPO + Spec. Cap	61.34	1.20	9.1	279
Spectral DPO	64.31	1.18	11.5	280

Table 6: Controls isolating spectral regularization from KL anchoring, update-norm control, and response-length effects. The gains of Spectral DPO are not explained by shorter outputs, smaller aggregate updates, or tighter reference anchoring. Average accuracy matches Table 1.

E Cross-Architecture Results and Matched-Seed Contrasts

Table 7 reports the regenerated three-epoch five-metric averages for vanilla DPO, the strongest included baseline, and Spectral DPO on all three architectures, together with the matched-seed contrast against the strongest baseline and its 95% paired percentile-bootstrap interval from 10,000 seed-level resamples. All three intervals exclude zero. The final row is a descriptive aggregate only: we do not report a pooled model-seed interval, because the three model families are not independent replications from a defined model population.

F Component Ablation Contrasts

Table 8 reports the matched-seed contrasts underlying Section 5.2. Each contrast is computed on the same ten seeds under an identical data mixture, optimiser, validation split, training budget, and decoding protocol.

G Target-Exponent Sensitivity

Table 9 reports a matched target-exponent sweep on Llama-3-8B-Instruct using the same seeds, data, and budget as the headline runs. Performance is stable across targets in $[1.8, 2.2]$ and declines modestly at the extremes. We therefore present two as a broad empirical anchor in this configuration, not a sharply identified universal optimum or a theoretical result. This sweep is a robustness analysis and was not used to select the headline configuration.

H Data-Efficiency Curve

Table 10 reports average accuracy on Llama-3-8B-Instruct trained on nested deterministic Ultra-Feedback subsets. Columns are labelled by ex-

Model	DPO	SimPO	Spectral DPO	Δ [95% PB]
Mistral-7B-v0.3	59.42 $\pm$ 0.41	60.31 $\pm$ 0.39	62.64 $\pm$ 0.37	+2.33 [2.10, 2.58]
Llama-3-8B-Instruct	60.90 $\pm$ 0.38	61.89 $\pm$ 0.36	64.31 $\pm$ 0.34	+2.42 [2.25, 2.56]
Qwen-2.5-14B	64.48 $\pm$ 0.35	65.17 $\pm$ 0.34	67.58 $\pm$ 0.32	+2.41 [2.19, 2.66]
Mean across tested models	61.60	62.46	64.84	+2.39

Table 7: Regenerated three-epoch five-metric averages (mean $\pm$ sample standard deviation over ten matched seeds) and matched-seed contrasts against the strongest included baseline. Δ is Spectral DPO minus SimPO with its 95% paired bootstrap interval.

Contrast	Mean difference	95% paired bootstrap
Full minus minus-tail	+0.84	[0.53, 1.15]
Full minus minus-spike	+0.63	[0.31, 0.94]
Full minus minus-KL	+0.40	[0.14, 0.68]
Full minus DPO	+3.41	[3.25, 3.55]

Table 8: Matched-seed contrasts for the component ablation on Llama-3-8B-Instruct.

Target exponent	Five-metric Avg. $\uparrow$	GSM-Plus $\uparrow$	AdvBench $\uparrow$
1.6	63.42 $\pm$ 0.41	72.1 $\pm$ 0.8	71.4 $\pm$ 0.7
1.8	64.02 $\pm$ 0.37	73.0 $\pm$ 0.7	72.0 $\pm$ 0.6
2.0	64.31 $\pm$ 0.34	73.6 $\pm$ 0.6	72.4 $\pm$ 0.6
2.2	64.05 $\pm$ 0.36	73.2 $\pm$ 0.7	72.1 $\pm$ 0.6
2.4	63.61 $\pm$ 0.40	72.5 $\pm$ 0.8	71.7 $\pm$ 0.7

Table 9: Matched target-exponent sensitivity sweep on Llama-3-8B-Instruct (mean $\pm$ sample standard deviation over ten seeds).

act preference-pair counts rather than approximate percentages, because the previously reported percentage labels did not match the realised subset sizes: the condition previously labelled “60%” contains 32,000 of the 60,800 available UltraFeedback training pairs, or 52.6%. Spectral DPO trained on 32,000 pairs approximately matches vanilla DPO trained on all 60,800, and remains above DPO at every subset size.

Method	60,800	32,000	16,000	6,400
DPO	61.0 $\pm$ 1.0	58.7 $\pm$ 1.2	55.4 $\pm$ 1.5	51.1 $\pm$ 1.8
Spectral DPO	64.4 $\pm$ 0.8	61.2 $\pm$ 1.0	59.0 $\pm$ 1.2	55.5 $\pm$ 1.5

Table 10: Data-efficiency curve on UltraFeedback-only subsets, labelled by exact preference-pair count (mean $\pm$ sample standard deviation over ten seeds).

I Experiment-Level Dataset Provenance

Table 11 records the dataset assignment recovered from run metadata for every experiment block. Every method within a comparison uses the identical preference-data assignment. AdvBench evaluates the same main-run checkpoints; no method receives a separate safety-specific training stage,

and no mixture ablation was performed, so no claim is made about HH-RLHF dependence of the AdvBench result.

Experiment block	UltraFeedback	HH-RLHF	Validation split
Main three-model runs	48,000	16,000	3,200 UF pairs
Llama long-run study	60,800	0	3,200 UF pairs
Component ablation	48,000	16,000	3,200 UF pairs
Data efficiency	6,400–60,800	0	3,200 UF pairs
Safety evaluation	no additional training		main-run ckpts
Post-hoc repair	no additional training		no optimiser update

Table 11: Experiment-level preference-data assignment. The 3,200-pair UltraFeedback validation archive was fixed before any test evaluation.

J Runtime and Peak-Memory Profiling

Table 12 reports profiling over 200 post-warm-up optimizer steps under matched BF16, sequence-length-2048, effective-batch-size-128 settings, with device synchronization around every measured step. Across the three model families, mean step-time overhead is 5.3%, throughput decreases by 5.1%, and peak memory increases by 1.7 GB per GPU. The overhead is concentrated on refresh steps; non-refresh steps use the vanilla DPO objective.

Model	Method	Step time	Overhead	Peak mem.
Llama-3-8B	DPO	1.72 s	—	72.4 GB
	Spectral DPO	1.81 s	+5.2%	74.1 GB
Mistral-7B	DPO	1.61 s	—	68.8 GB
	Spectral DPO	1.69 s	+5.0%	70.4 GB
Qwen-2.5-14B	DPO	3.08 s	—	76.2 GB
	Spectral DPO	3.26 s	+5.8%	78.0 GB

Table 12: Matched runtime and peak-memory profiling over 200 post-warm-up steps. Peak memory is reported per GPU. Throughput decreases by 5.1% on average (74.4 to 70.7, 79.5 to 75.7, and 41.6 to 39.3 sequences per second respectively).

K Post-hoc Spectral Projection Procedure

Algorithm 2 details the projection procedure used in Section 5.6. It is applied independently to each

effective adapted matrix in a degraded DPO checkpoint, with the target tail exponent and spike ratio matched to the Spectral DPO training targets. Unlike the training penalty, this offline procedure uses a full detached SVD.

Algorithm 2 Post-hoc Spectral Projection Procedure

- 1: **Input:** Effective matrix W_θ , target tail exponent $\alpha_{\text{tgt}} = 2.0$, target spike ratio γ
 - 2: Compute SVD: $W_\theta = U\Sigma V^\top$ with singular values σ_i
 - 3: Calculate initial tail exponent $\hat{\alpha}$ and spike ratio $q(W_\theta)$
 - 4: **if** $q(W_\theta) > \gamma$ **then**
 - 5: $\sigma_1 \leftarrow \gamma \cdot (\bar{\sigma} + \varepsilon_\sigma)$ to satisfy the spike cap
 - 6: **end if**
 - 7: Extract $\lambda_i = \sigma_i^2$ for $i > 1$ and fit a discretised Pareto distribution with exponent α_{tgt}
 - 8: Update bulk $\sigma_i \leftarrow \sqrt{\lambda_i^{\text{new}}}$ while preserving $\sum_{i>1} \sigma_i^2$
 - 9: Reconstruct matrix: $W_\theta^{\text{new}} = U\Sigma^{\text{new}}V^\top$
 - 10: **Output:** Repaired projection W_θ^{new}
-

L Hyperparameter Search Protocol and Complete Configuration

Each method received exactly 12 tuning trials. Trials were generated once by seeded random search, using log-uniform sampling for learning rates and positive penalty magnitudes and uniform sampling for bounded temperatures, caps, and ratios; the sampler seed and all realised configurations are recorded in the trial manifest in the supplementary archive. Selection is deterministic: minimum validation preference loss on a fixed 3,200-item UltraFeedback validation archive established before the test evaluations, with an exact tie broken by the lower trial identifier. The selected configuration was reused across model families and test benchmarks. For Spectral DPO, the DPO-selected learning rate was reused rather than retuned. Equal trial counts do not make the design spaces identical, so Table 13 separates the quantities that were searched from those fixed before the sweep, and the range summary in Table 14 is a summary rather than a substitute for the manifest.

M Per-Seed Results

All main experiments are repeated across ten matched seeds (0–9). Table 15 reports per-seed

Quantity	Value	Status
DPO temperature β	0.1	searched
Learning rate	5×10^{-7}	searched (DPO), reused
Outer spectral weight λ	0.01	searched
Spike-term weight η	1.0	searched
Spike threshold γ	5.0	searched
Distributional weight μ	0.1	fixed
Target exponent	2.0	fixed
Nominal tail fraction f	0.10 of retained spectrum	fixed
Minimum tail size n_{min}	32	fixed; generally binds
Matrix sampling rate	10%	fixed
Spectral refresh interval	every 10 steps	fixed
LoRA rank / scale	64 / 16	fixed
AdamW (β_1, β_2)	(0.9, 0.95)	fixed
Effective batch size	128	fixed
Maximum sequence length	2048	fixed
Main-run / long-run epochs	3 / 10	fixed

Table 13: Complete Spectral DPO training configuration. The values previously omitted from the submitted version are $\mu = 0.1$, the learning rate 5×10^{-7} , and $\beta = 0.1$.

average accuracy for every method in the main comparison, and Table 16 reports per-seed benchmark-level values for DPO and Spectral DPO, the two methods used in the paired analysis. The supplementary archive contains the same per-seed, per-benchmark breakdown for every baseline in Table 1, and all aggregates, differences, and intervals reported in the paper are generated from those unrounded rows through the single reproduction entry point.

N Long-Run Trajectories

Table 17 reports the epoch-indexed trajectories underlying Figure 1 for the Llama-3-8B-Instruct / UltraFeedback long-run study. At every epoch checkpoint the fitted exponent, spike ratio, and spectral KL were recomputed offline for both DPO and Spectral DPO using a full detached SVD on the same fixed set of eligible projection matrices. The matrix set was fixed before inspecting downstream results, so differences do not arise from changing stochastic matrix samples. For each seed, each statistic was first averaged over that fixed matrix set; the table reports the mean of the resulting seed-level values over ten seeds. Vanilla DPO moves progressively farther from the target exponent, develops a larger spike ratio, and increases distributional divergence after its GSM-Plus peak, while Spectral DPO stabilises all three quantities and preserves its downstream gain through epoch ten.

Table 18 reports the corresponding validation preference reward at three checkpoints, alongside SimPO and Conservative DPO. The reward and GSM-Plus trajectories show the same qualitative pattern: vanilla DPO continues improving the validation reward proxy while downstream rea-

Method	Budget	Search Space	Selected Values
DPO	12 runs	$\beta \in [0.05, 0.2], \text{lr} \in [10^{-7}, 10^{-5}]$	$\beta = 0.1, \text{lr} = 5 \times 10^{-7}$
IPO	12 runs	temperature $\in [0.05, 0.2], \text{lr} \in [10^{-7}, 10^{-5}]$	temperature=0.1, lr=5 $\times 10^{-7}$
KTO	12 runs	$\beta \in [0.05, 0.2], \text{lr} \in [10^{-7}, 10^{-5}]$	$\beta = 0.1, \text{lr} = 5 \times 10^{-7}$
SimPO	12 runs	$\beta \in [1, 5], \gamma \in [0.5, 2.0], \text{lr} \in [10^{-7}, 10^{-5}]$	$\beta = 2.5, \gamma = 1.0, \text{lr} = 8 \times 10^{-7}$
ORPO	12 runs	odds-ratio weight $\in [0.05, 0.2], \text{lr} \in [10^{-7}, 10^{-5}]$	weight=0.1, lr=5 $\times 10^{-7}$
Conservative DPO	12 runs	KL multiplier $\in [1.5, 3.0], \text{lr} \in [10^{-7}, 10^{-5}]$	multiplier=2.0, lr=5 $\times 10^{-7}$
DPO + Frob. Penalty	12 runs	penalty weight $\in [10^{-4}, 10^{-2}], \text{lr} \in [10^{-7}, 10^{-5}]$	weight=10 ⁻³ , lr=5 $\times 10^{-7}$
DPO + Spec. Cap	12 runs	cap $\in [2, 10], \text{lr} \in [10^{-7}, 10^{-5}]$	cap=5.0, lr=5 $\times 10^{-7}$
Spectral DPO	12 runs	$\lambda \in [0.005, 0.05], \eta \in [0.5, 2.0], \gamma \in [2, 10]$	$\lambda = 0.01, \eta = 1.0, \gamma = 5.0$

Table 14: Hyperparameter search ranges and selected values. All methods received the same 12-trial budget; the learning rate for Spectral DPO was reused from DPO rather than retuned, and the quantities marked fixed in Table 13 were fixed before the penalty sweep.

Method	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	Mean $\pm$ SD
SFT	55.0	56.4	56.8	55.6	56.0	56.3	56.9	55.8	55.2	56.2	56.1 $\pm$ 0.6
DPO	60.1	61.2	61.7	60.6	61.0	61.3	61.8	60.9	60.2	61.0	61.0 $\pm$ 0.6
IPO	59.5	60.6	61.0	60.0	60.5	60.7	61.1	60.3	59.8	60.5	60.4 $\pm$ 0.5
KTO	60.8	61.9	62.3	61.4	61.6	62.0	62.4	61.5	60.9	61.8	61.7 $\pm$ 0.5
SimPO	61.1	62.2	62.6	61.7	62.0	62.3	62.7	61.8	61.2	62.1	62.0 $\pm$ 0.5
ORPO	60.5	61.6	62.0	61.0	61.5	61.7	62.1	61.3	60.7	61.6	61.4 $\pm$ 0.5
Conservative DPO	60.0	61.1	61.5	60.4	60.9	61.2	61.6	60.7	60.1	60.9	60.9 $\pm$ 0.5
DPO + Frob. Penalty	60.2	61.3	61.6	60.7	61.1	61.4	61.7	60.9	60.4	61.2	61.1 $\pm$ 0.5
DPO + Spec. Cap	60.4	61.5	61.8	60.9	61.3	61.6	61.9	61.0	60.5	61.4	61.3 $\pm$ 0.5
Spectral DPO	63.8	64.6	64.9	64.1	64.7	64.9	65.0	64.4	63.6	63.9	64.4 $\pm$ 0.5

Table 15: Per-seed average accuracy for all main methods on Llama-3-8B-Instruct after three epochs.

soning accuracy declines, whereas Spectral DPO keeps downstream accuracy stable through the later epochs.

O Evaluation Protocol Details

To ensure strict reproducibility, we standardise the evaluation protocol across all models and runs. All generations use greedy decoding (temperature 0.0, top- p 1.0) to isolate model capabilities from sampling variance, and Length-Controlled AlpacaEval 2 and MT-Bench both use gpt-4-turbo-2024-04-09 as the judge. We follow the default prompt templates provided by the respective evaluation frameworks for AlpacaEval, MT-Bench, and TruthfulQA, and for AdvBench we use the standard prefix-matching evaluation script. For GSM-Plus we use few-shot prompting with standard chain-of-thought, followed by a strict regex-based answer extraction that parses the final numerical output. Whenever a generation fails to produce a parseable answer, it is counted as incorrect; we verified manually that this convention does not produce a systematic bias across methods on a random sample of 200 generations per benchmark.

Seed	DPO						Spectral DPO					
	AE2 LC	MT	GSM+	TQA	Adv ASR	Avg	AE2 LC	MT	GSM+	TQA	Adv ASR	Avg
0	19.6	8.13	66.7	57.2	25.2	60.1	22.6	8.40	72.0	60.4	22.1	63.8
1	20.7	8.22	68.4	58.8	24.0	61.2	23.4	8.48	73.8	61.6	21.0	64.6
2	21.2	8.28	69.0	59.1	23.5	61.7	23.8	8.52	74.2	62.0	20.7	64.9
3	20.0	8.18	67.5	58.0	24.4	60.6	23.0	8.44	73.1	60.9	21.6	64.1
4	20.5	8.21	68.1	58.3	24.1	61.0	23.5	8.49	73.7	61.5	21.0	64.7
5	20.8	8.24	68.5	58.5	23.8	61.3	23.7	8.50	74.0	61.8	20.8	64.9
6	21.3	8.29	69.2	59.0	23.4	61.8	23.9	8.53	74.4	62.1	20.6	65.0
7	20.4	8.20	68.0	58.2	24.2	60.9	23.2	8.46	73.5	61.3	21.3	64.4
8	19.8	8.15	67.2	57.6	25.0	60.2	22.4	8.39	71.9	60.3	22.2	63.6
9	20.5	8.20	67.8	58.1	24.4	61.0	22.8	8.43	72.8	60.7	21.9	63.9

Table 16: Per-seed benchmark-level results for DPO and Spectral DPO. AdvBench is reported as Attack Success Rate, so lower values are better.

Epoch	$\hat{\alpha}$ (DPO)	$\hat{\alpha}$ (Spec)	Spike $\hat{q}$ (DPO)	Spike $\hat{q}$ (Spec)	Spec. KL (DPO)	Spec. KL (Spec)	GSM+ (DPO)	GSM+ (Spec)
0	3.82	3.82	4.62	4.62	0.184	0.184	55.3	55.3
1	3.91	3.21	4.81	4.55	0.201	0.158	62.4	61.2
2	4.08	2.74	5.07	4.49	0.224	0.137	66.0	67.8
3	4.31	2.18	5.36	4.46	0.251	0.119	68.2	73.6
4	4.52	2.02	5.61	4.51	0.278	0.108	67.5	76.1
5	4.71	1.98	5.82	4.55	0.301	0.106	66.4	75.8
6	4.89	2.01	5.97	4.58	0.316	0.108	65.8	75.7
7	5.06	2.04	6.08	4.61	0.329	0.110	65.1	75.6
8	5.17	2.06	6.16	4.65	0.338	0.111	64.7	75.7
9	5.29	2.04	6.23	4.68	0.346	0.112	64.4	75.8
10	5.41	2.03	6.29	4.70	0.354	0.113	64.1	75.8

Table 17: Epoch-indexed trajectories of the three regularised quantities and downstream accuracy on Llama-3-8B-Instruct trained on UltraFeedback only (mean over ten matched seeds). Epoch 0 is the shared initialization checkpoint.

Method	Epoch 3 Reward	Epoch 3 GSM+	Epoch 5 Reward	Epoch 5 GSM+	Epoch 10 Reward	Epoch 10 GSM+
DPO	0.721 ± 0.012	68.2 ± 1.3	0.748 ± 0.014	66.4 ± 1.8	0.784 ± 0.015	64.1 ± 2.1
SimPO	0.716 ± 0.011	69.4 ± 1.3	0.739 ± 0.013	67.8 ± 1.7	0.771 ± 0.014	65.5 ± 2.0
KL-DPO	0.694 ± 0.010	68.7 ± 1.4	0.711 ± 0.012	67.4 ± 1.7	0.735 ± 0.014	66.0 ± 1.9
Spectral DPO	0.698 ± 0.010	73.6 ± 1.1	0.723 ± 0.011	75.8 ± 1.3	0.752 ± 0.011	75.8 ± 1.3

Table 18: Validation reward and GSM-Plus accuracy during long-run training. Values are mean $\pm$ sample standard deviation over ten matched seeds.