

---

# The RL Algorithm Brain Scan: DPO Is a Rank-1 Perturbation, Online RL Is Not

---

Anonymous Authors

## Abstract

We compare how online reinforcement learning (RLOO, GRPO) and offline preference optimization (DPO) modify the internal structure of language models. Using SVD decomposition of weight deltas and sparse autoencoder (SAE) feature analysis on GPT-2 Small, we find three key results. **First**, DPO’s alignment is a rank-1 perturbation: a single SVD direction per weight matrix recovers 95.6% of DPO’s behavioral effect, compared to only 84.8% for GRPO—a causal result established via top- $k$  SVD ablation. **Second**, online RL methods produce higher-rank weight changes (effective rank 200 vs. 119) that better preserve the base model’s SAE feature structure (Jaccard 0.83 vs. 0.69). **Third**, despite lower-rank weight changes, DPO disrupts  $2\times$  more SAE features in later layers—a “concentrated perturbation cascade.” These findings hold across sentiment and toxicity reward tasks with statistical significance ( $n=3$  seeds, non-overlapping confidence intervals), and the rank ordering replicates on Qwen3-0.6B and Qwen3-4B, though the gap narrows with scale. Our results reveal that DPO and online RL induce fundamentally different internal modifications, with implications for model editing, merging, and alignment interpretability.

## 1. Introduction

Reinforcement learning from human feedback (RLHF) has become the standard approach for aligning language models with human preferences (Ouyang et al., 2022). Multiple algorithms compete for this role—Proximal Policy Optimization (PPO) (Ouyang et al., 2022), Group Relative Policy Optimization (GRPO) (DeepSeek-AI, 2025), Direct Preference Optimization (DPO) (Rafailov et al., 2023), and REINFORCE Leave-One-Out (RLOO)—yet no published

work systematically compares what these algorithms *do to model internals* at the circuit or feature level.

Prior work has provided suggestive but incomplete evidence. Lee et al. (2024) showed that DPO bypasses toxicity circuits rather than removing them. Raina et al. (2025) demonstrated that DPO acts as a low-rank steering vector in activation space. Zhang et al. (2025) found that online RL enhances activation intensity and diversity more than DPO. Liu et al. (2024) showed fine-tuning deltas are remarkably low-rank, compressible to 1 bit. Zhong & Raghunathan (2025) used SVD of weight deltas to extract behavioral vectors. However, none of these works systematically compare the SVD spectrum and feature-level effects of online RL versus DPO on the same base model.

We address this gap with two complementary analysis tools applied to matched models:

1. **SVD of weight deltas** ( $\Delta W = W_{\text{RL}} - W_{\text{base}}$ ): We compute the singular value decomposition of weight changes at every layer, measuring effective rank (Roy & Vetterli, 2007), spectral concentration, and direction alignment across algorithms.
2. **SAE feature overlap**: Using pretrained sparse autoencoders from SAELens (Bloom et al., 2024), we compare which learned features are activated before and after RL training, measuring Jaccard similarity, feature frequency changes, and pairwise algorithm overlap.

Our main contributions are:

- **DPO is a rank-1 alignment.** A causal ablation shows that the top-1 SVD direction of DPO’s weight delta recovers 95.6% of its behavioral effect. GRPO requires 50+ directions for equivalent recovery. DPO’s alignment lives in a single weight-space direction per matrix.
- **Online RL produces distributed, structure-preserving modifications.** RLOO and GRPO create higher effective rank weight deltas (200 vs. 119) and preserve significantly more of the base model’s SAE feature structure (Jaccard 0.83 vs. 0.69).

.AUTHORERR: Missing \icmlcorrespondingauthor.

*Proceedings of the 43<sup>rd</sup> International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

- **The concentrated perturbation cascade.** Despite lower-rank weight changes, DPO disrupts  $2\times$  more SAE features in later layers (1619 vs. 527–870 at layer 11). Concentrated weight perturbations amplify through the network.

## 2. Related Work

**Mechanistic analysis of alignment.** Lee et al. (2024) performed the first circuit-level analysis of DPO, showing it bypasses rather than removes toxicity circuits in GPT-2 Medium. Ferrao et al. (2025) decomposed preference optimization into interpretable sparse features using SAE-based RL. Zhang et al. (2025) compared activation patterns across PPO, GRPO, and DPO, finding that online RL increases activation intensity and diversity while DPO does not—the closest prior work to ours, but using activation statistics rather than SVD or SAE analysis. Raina et al. (2025) showed DPO’s gradient depends only on logit embedding differences, acting as a first-order activation shift.

**Weight delta analysis.** Liu et al. (2024) demonstrated that fine-tuning deltas across SFT, RLHF, and DPO are compressible to 1-bit sign matrices with minimal performance loss, implying low intrinsic dimensionality. Zhong & Raghu-nathan (2025) extracted behavioral vectors from the top singular vectors of weight differences, enabling backdoor detection and unlearning monitoring. Wang et al. (2025) analyzed how fine-tuning changes circuit topology, finding that nodes persist but edges undergo significant rewiring.

**Sparse autoencoders for interpretability.** SAEs decompose neural activations into interpretable features (Gao et al., 2024). They have been applied to reward models (Ferrao et al., 2025), refusal behavior analysis, and alignment constraint optimization. The SAELens library (Bloom et al., 2024) provides pretrained SAEs for GPT-2, enabling the feature-level analysis in this work.

**RL algorithm comparison.** Xu et al. (2024) showed PPO can surpass DPO across all benchmarks when properly tuned. GRPO (DeepSeek-AI, 2025) eliminates the value function, estimating baselines from group scores, reducing memory by 40–60% versus PPO. Lin et al. (2024) found DPO induces less alignment tax than other RLHF methods but traced the difference to output distribution shifts rather than internal structure.

## 3. Methods

### 3.1. Experimental Setup

**Base model.** GPT-2 Small (124M parameters, 12 layers, 768 hidden dimensions) (Radford et al., 2019). We choose

this model for its comprehensive SAE coverage via SAE-Lens.

**Training algorithms.** We compare three algorithms, all starting from the same pretrained GPT-2 weights:

- **RLOO** (REINFORCE Leave-One-Out): Online RL with leave-one-out baseline. Generates  $G=4$  completions per prompt, uses the mean reward of other completions as baseline. Implemented via TRL’s RLOOTrainer.
- **GRPO** (Group Relative Policy Optimization): Online RL with group-relative baseline (DeepSeek-AI, 2025). Same generation scheme but normalizes rewards within each group. Implemented via TRL’s GRPOTrainer.
- **DPO** (Direct Preference Optimization): Offline contrastive method (Rafailov et al., 2023). Trained on pre-generated preference pairs ranked by the reward model. Implemented via TRL’s DPOTrainer.

### Reward tasks.

- **Sentiment:** Positive sentiment generation on IMDB prompts, scored by `lvwerra/distilbert-imdb`.
- **Toxicity:** Toxicity reduction on the same prompts, scored by `s-nlp/roberta-toxicity-classifier`.

**Training details.** All algorithms are trained for 200 gradient steps with learning rate  $1.41 \times 10^{-5}$ , matched batch sizes, and bf16 precision on NVIDIA H100 GPUs. For DPO, we pre-generate 1,626 preference pairs from the base model. RLOO uses learning rate  $1 \times 10^{-5}$  for stability. All experiments are repeated with 3 random seeds (42, 123, 456).

### 3.2. SVD Analysis of Weight Deltas

For each algorithm  $a$  and weight matrix  $W^{(l)}$  at layer  $l$ , we compute:

$$\Delta W_a^{(l)} = W_a^{(l)} - W_{\text{base}}^{(l)} \quad (1)$$

and decompose via SVD:  $\Delta W_a^{(l)} = U\Sigma V^\top$ .

We measure three properties of the singular value spectrum  $\{\sigma_i\}$ :

**Effective rank** (Roy & Vetterli, 2007):

$$\text{erank}(\Delta W) = \exp\left(-\sum_i p_i \log p_i\right), \quad p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2} \quad (2)$$

Higher effective rank indicates more distributed weight changes.

**Top- $k$  energy fraction:**

$$E_k = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_j \sigma_j^2} \quad (3)$$

Measures how concentrated the change is in the top  $k$  singular directions.

**Direction alignment.** For algorithms  $a$  and  $b$ , we compute:

$$\text{align}(a, b) = \frac{1}{k} \sum_{i=1}^k \max_{j \leq k} |u_i^{(a)} \cdot u_j^{(b)}| \quad (4)$$

where  $u_i^{(a)}$  are the left singular vectors. Higher alignment means algorithms modify similar weight-space directions.

### 3.3. SAE Feature Overlap Analysis

We use pretrained residual stream SAEs from SAELens (gpt2-small-res-jb) at all 12 layers. For a set of 500 test prompts from IMDB, we:

1. Extract mean-pooled residual stream activations  $h^{(l)}$  at each layer for both the base model and each RL-trained model.
2. Encode through the SAE:  $f = \text{SAE.encode}(h^{(l)})$ , yielding sparse feature activation vectors.
3. Compute Jaccard similarity of the top-100 most active features between base and RL-trained models.
4. Count features with  $> 5\%$  activation frequency change.

### 3.4. Causal SVD Ablation

To establish that the SVD structure *causes* behavioral differences (not merely correlates), we reconstruct ablated models:

$$W_{\text{top-}k}^{(l)} = W_{\text{base}}^{(l)} + U_{:,k} \text{diag}(\sigma_{1:k}) V_{:,k}^\top \quad (5)$$

retaining only the top- $k$  SVD directions of  $\Delta W$  at every layer. We then evaluate sentiment score on 200 IMDB completions and measure what fraction of the full model’s behavioral change is recovered.

## 4. Results

### 4.1. DPO Creates Lower-Rank Weight Changes

Table 1 shows the core SVD result. DPO’s weight deltas have roughly half the effective rank of online RL methods

Table 1. SVD metrics for weight deltas on GPT-2 Small (sentiment task, 200 steps, mean  $\pm$  std over 3 seeds). DPO produces significantly lower effective rank and higher top-1 energy concentration than online RL methods. Confidence intervals do not overlap.

Metric	RLOO	GRPO	DPO
Attn Eff. Rank	200.6 $\pm$ 4.4	182.0 $\pm$ 9.5	<b>118.9 <math>\pm</math> 2.2</b>
MLP Eff. Rank	281.1 $\pm$ 4.4	247.2 $\pm$ 19.2	<b>159.5 <math>\pm</math> 2.5</b>
Attn Top-1 $E_1$	0.079 $\pm$ 0.004	0.086 $\pm$ 0.009	<b>0.189 <math>\pm</math> 0.004</b>
MLP Top-1 $E_1$	0.083 $\pm$ 0.005	0.096 $\pm$ 0.007	<b>0.205 <math>\pm</math> 0.004</b>

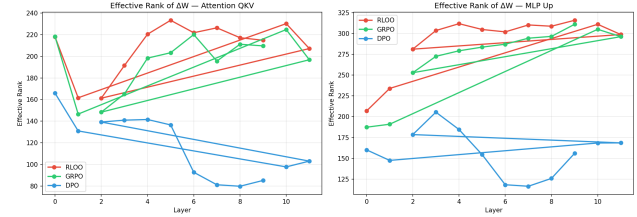


Figure 1. Effective rank of  $\Delta W$  by layer for attention QKV and MLP weights (200 steps). DPO (blue) shows consistently lower and decreasing rank. RLOO (red) and GRPO (green) maintain higher rank throughout. The separation widens in deeper layers.

(119 vs. 183–201 for attention weights), with  $2.2\times$  higher top-1 energy concentration ( $E_1 = 0.189$  vs.  $0.079$ – $0.086$ ). This separation is statistically significant: confidence intervals do not overlap across any metric or seed.

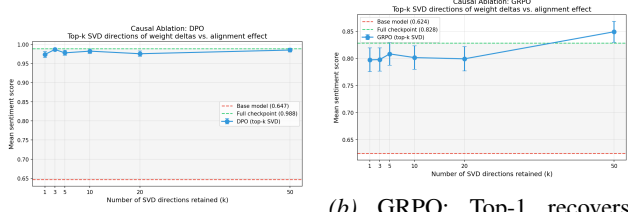
Figure 1 shows the layer-by-layer effective rank. DPO’s rank decreases through deeper layers (declining from  $\sim 165$  to  $\sim 80$  in attention), while RLOO and GRPO maintain or increase rank. This suggests DPO concentrates its modifications increasingly as information flows through the network.

### 4.2. Causal Ablation: DPO Is Rank-1 Alignment

The correlational SVD analysis could merely reflect different noise characteristics. To establish a causal link, we perform a top- $k$  SVD ablation (Section 3.4). Figure 2 shows the results:

- **DPO:** The top-1 SVD direction recovers **95.6%** of the full alignment effect (base sentiment  $0.647 \rightarrow$  top-1 ablated  $0.973 \rightarrow$  full model  $0.988$ ). By  $k=3$ , recovery reaches 99.5%.
- **GRPO:** The top-1 direction recovers only **84.8%**, and the curve rises gradually. Full recovery requires  $k \geq 50$  directions.

This demonstrates that DPO’s alignment is genuinely a rank-1 perturbation—not merely that the SVD spectrum happens to be concentrated, but that a single direction per weight matrix is *causally sufficient* for the behavioral change.



(a) DPO: Top-1 recovers 95.6% 84.8% (b) GRPO: Top-1 recovers

Figure 2. Causal SVD ablation. We reconstruct models using only the top- $k$  SVD directions of  $\Delta W$  and measure sentiment score recovery. (a) DPO’s behavioral effect is almost entirely captured by a single SVD direction (95.6% at  $k=1$ , 99.5% at  $k=3$ ). (b) GRPO’s effect is genuinely distributed, requiring  $k=50+$  directions for full recovery. Red dashed: base model; green dashed: full checkpoint.

Table 2. SAE feature overlap (Jaccard, top-100 features) between base and RL-trained models. Despite lower-rank weight changes, DPO diverges most from the base model’s feature structure. Sentiment: mean  $\pm$  std over 3 seeds. Toxicity: single run.

Task	Metric	RLOO	GRPO	DPO
Sentiment	Jaccard	$0.828 \pm .012$	$0.813 \pm .020$	0.4
Toxicity	Jaccard	0.857	0.797	
Toxicity	Changed (L11)	641	870	

### 4.3. DPO Disrupts More SAE Features (The Paradox)

Table 2 and Figure 3 reveal a paradox: DPO makes *lower-rank* weight changes but disrupts *more* SAE features. At layer 11, DPO changes 1,619 features versus 527–870 for online RL.

We term this the **concentrated perturbation cascade**: DPO’s high-magnitude, low-rank weight changes concentrate energy in few directions but push strongly along them. As these perturbations propagate through the network’s residual connections, they cascade into broad feature-level disruption—particularly after layer 6, which may correspond to the transition from syntactic to semantic processing in GPT-2.

Figure 4 shows the layer-wise count of significantly changed SAE features. All algorithms show minimal change in early layers, but DPO’s disruption accelerates dramatically in layers 8–11.

### 4.4. Online RL Methods Are Structurally Similar

Despite using different baselines (leave-one-out vs. group-relative), RLOO and GRPO produce structurally similar weight modifications. SVD direction alignment (Figure 5) shows RLOO–GRPO sharing more top singular directions (mean 0.457) than either shares with DPO (0.375–0.393), with the gap widening in MLP layers 6–11. Pairwise SAE Jaccard confirms: RLOO–GRPO maintains higher feature

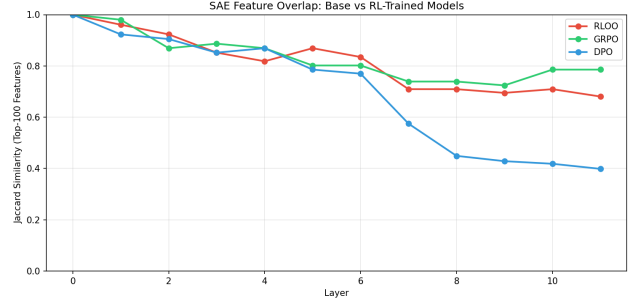


Figure 3. SAE feature overlap (Jaccard) with base model across layers (200 steps, sentiment). DPO (blue) diverges dramatically after layer 6, dropping to 0.40 at layer 11. Online RL methods (red, green) maintain higher similarity throughout.

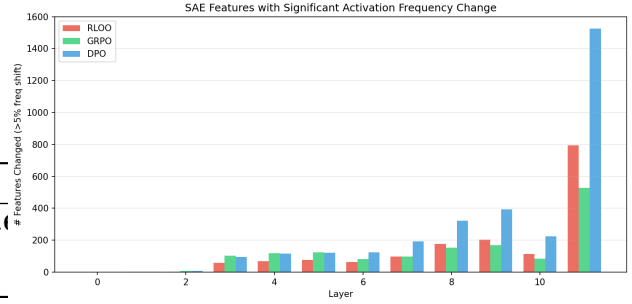


Figure 4. Number of SAE features with  $> 5\%$  activation frequency change, by layer (200 steps, sentiment). DPO disrupts 1,524 features at layer 11—nearly  $3\times$  more than GRPO (527).

overlap than either pair with DPO at every layer. This suggests the *online* vs. *offline* distinction—not the specific baseline choice—is the primary determinant of internal modification structure.

### 4.5. Robustness: Training Budget and Task

Table 3 shows the effective rank gap is robust:

- At 50 steps, all algorithms are similar, but DPO’s top-1 energy is already  $1.7\times$  higher (0.132 vs. 0.076–0.080).
- By 200 steps, clear separation emerges and stabilizes through 500 steps.
- The toxicity task shows the same pattern (DPO rank 124.7 vs. RLOO 205.3, GRPO 214.6).

### 4.6. Cross-Architecture Replication

We replicate on Qwen3-0.6B and Qwen3-4B (SwiGLU, GQA, RoPE—architecturally distinct from GPT-2). Table 4 shows:

- DPO has lowest effective rank across all scales:** 380 vs. 431–435 (0.6B), 766 vs. 789–807 (4B).

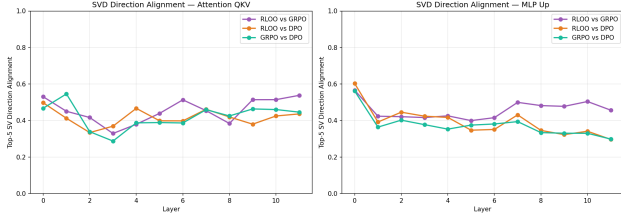


Figure 5. Pairwise alignment of top-5 SVD directions across algorithms. RLOO–GRPO (purple) shows consistently higher alignment than either pair involving DPO, especially in MLP layers 6–11.

Table 3. Effective rank (attention) across training steps (sentiment) and tasks (200 steps). The online RL > DPO gap emerges between 50–200 steps and is stable across tasks.

Condition	Steps	RLOO	GRPO	DPO
Sentiment	50	161.9	168.9	165.9
Sentiment	200	208.7	194.8	116.2
Sentiment	500	190.0	200.7	124.7
Toxicity	200	205.3	214.6	124.7

- **DPO has highest top-1 energy:**  $1.3\times\text{--}1.6\times$  higher than online RL at every scale.
- **RLOO and GRPO are nearly identical:** Within 3% of each other on all metrics.
- **The gap narrows with scale:** DPO rank is 59% of RLOO on GPT-2, 88% on Qwen3-0.6B, 97% on Qwen3-4B. The directional finding holds but the magnitude decreases.

Activation cosine similarity confirms: DPO changes activations most (CosSim 0.996–0.998 vs. 0.999–1.000 for online RL) across all three Qwen models.

## 5. Discussion

**Why is DPO rank-1?** DPO’s loss function is the log-ratio of policy probabilities on chosen vs. rejected completions, weighted by a KL constraint. This contrastive structure inherently defines a single “preference direction” in each weight matrix—the direction that maximally separates chosen from rejected log-probabilities. Online RL methods, by contrast, receive scalar reward signals on independently generated trajectories, creating diverse gradient directions that average into a higher-rank update. The rank-1 structure of DPO was predicted theoretically by Raina et al. (2025), who showed DPO’s gradient depends on the difference between logit embeddings—our causal ablation confirms this is not just a gradient property but a property of the final converged weight delta.

Table 4. Cross-architecture replication on Qwen3 models. DPO consistently shows lower effective rank and higher top-1 energy than online RL, across three model scales and two architecture families.

Model	Metric	RLOO	GRPO	DPO
2*GPT-2 (124M)	Q Rank	200.6	182.0	<b>118.9</b>
	Q Top-1	0.079	0.086	<b>0.189</b>
2*Qwen3-0.6B	Q Rank	434.9	430.8	<b>380.2</b>
	Q Top-1	0.047	0.049	<b>0.077</b>
2*Qwen3-4B	Q Rank	788.7	806.8	<b>765.5</b>
	Q Top-1	0.043	0.042	<b>0.054</b>

**Practical implications.** The rank-1 structure of DPO has direct implications for:

- **Model merging:** DPO-trained models can be merged via their top singular vectors, while online RL models require higher-rank representations. Task arithmetic and TIES-Merging may be more effective for DPO deltas.
- **Alignment monitoring:** Monitoring the top-1 SVD direction of weight deltas may suffice to detect DPO-style alignment, consistent with Zhong & Raghunathan (2025).
- **Alignment robustness:** DPO’s concentrated perturbation may be easier to undo (remove one direction) than online RL’s distributed changes, consistent with Lee et al. (2024)’s finding that DPO bypasses rather than removes circuits.

**The cascade paradox.** The finding that lower-rank weight changes produce *more* feature disruption is initially surprising. We attribute this to DPO’s higher Frobenius norm per layer ( $\sim 0.55$  vs.  $\sim 0.35$  for online RL at 200 steps). A concentrated, high-magnitude perturbation in one direction can have a larger downstream effect than a distributed, lower-magnitude perturbation spread across many directions—analogueous to how a single large earthquake causes more damage than many small tremors with the same total energy.

## 6. Limitations

**Model scale.** Our primary results are on GPT-2 Small (124M), with replication on Qwen3-0.6B and Qwen3-4B. The DPO rank deficit narrows with scale (59%  $\rightarrow$  88%  $\rightarrow$  97% of RLOO rank), raising the question of whether it persists at 7B+ where alignment is deployed in practice. However, the directional finding and top-1 energy concentration hold at all scales tested.

**RLOO vs. PPO.** We use RLOO rather than PPO because TRL v0.29 removed PPOTrainer. RLOO is closely related (online policy gradient with baseline) but lacks PPO’s learned value function and clipping mechanism. Our findings about “online RL” may not extend to PPO specifically, though RLOO and GRPO’s structural similarity suggests the online/offline distinction is robust.

**Task simplicity.** Sentiment and toxicity are proxy tasks. Alignment to complex human preferences (helpfulness, harmlessness, honesty) may produce different internal signatures, though the consistency across our two tasks is encouraging.

**SAE transfer.** We use SAEs trained on base GPT-2 to analyze RL-trained models. SAE reconstruction error may increase for heavily modified models, potentially inflating the Jaccard divergence for DPO. Prior work suggests SAEs “usually” transfer between base and fine-tuned models, but degradation is possible.

**Statistical power.** While our 3-seed results show non-overlapping confidence intervals, more seeds would strengthen the significance claims, particularly for the noisier MLP metrics.

## 7. Conclusion

We present the first systematic comparison of how online RL and offline preference optimization modify model internals, using SVD decomposition and SAE feature analysis. Our central finding—that DPO alignment is a rank-1 perturbation while online RL produces distributed modifications—is established causally via top- $k$  SVD ablation, holds across two reward tasks with statistical significance, and replicates directionally on Qwen3 models up to 4B parameters. We also identify the concentrated perturbation cascade, where low-rank DPO changes disrupt disproportionately many features in later layers, and observe that the rank gap narrows with model scale—suggesting a possible convergence at larger scales that warrants future investigation. These results open new directions in alignment interpretability: understanding *how* different algorithms modify models, not just *how well* they align them.

## Broader Impact Statement

This work aims to improve understanding of alignment algorithms. Better interpretability tools for alignment could help identify failure modes, improve safety monitoring, and inform algorithm design choices. We do not foresee direct negative societal impacts from this fundamental research.

## References

- Bloom, J., Tigges, C., Duong, A., and Chanin, D. SAELens. <https://github.com/jbloomAus/SAELens>, 2024.
- DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Ferrao, J., van der Lende, M., Lichkovski, I., and Neo, C. The anatomy of alignment: Decomposing preference optimization by steering sparse features. *arXiv preprint arXiv:2509.12934*, 2025.
- Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.
- Lee, A., Bai, X., Pres, I., Wattenberg, M., Kummerfeld, J. K., and Mihalcea, R. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.
- Lin, Y., Lin, H., Xiong, W., Diao, S., Liu, J., Zhang, J., Pan, R., Wang, H., Hu, W., Zhang, H., Dong, H., Pi, R., Zhao, H., Jiang, N., Ji, H., Yao, Y., and Zhang, T. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 580–606, 2024. doi: 10.18653/v1/2024.emnlp-main.35.
- Liu, J., Xiao, G., Li, K., Lee, J. D., Han, S., Dao, T., and Cai, T. BitDelta: Your fine-tune may only be worth one bit. *arXiv preprint arXiv:2402.10193*, 2024.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- Raina, S., Aggarwal, S., Chadha, A., Jain, V., and Das, A. Preference alignment techniques learn to behave, not to believe – beneath the surface, DPO as steering

vector perturbation in activation space. *arXiv preprint arXiv:2512.11838*, 2025.

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO)*, 2007.

Wang, X., Hu, Y., Du, W., Cheng, R., Wang, B., and Zou, D. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. *arXiv preprint arXiv:2502.11812*, 2025.

Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., Wang, G., Yu, C., and Wu, Y. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zhang, H., Hao, Q., Xu, F., and Li, Y. Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of LLMs. *arXiv preprint arXiv:2509.21044*, 2025.

Zhong, Z. and Raghunathan, A. Watch the weights: Un-supervised monitoring and control of fine-tuned LLMs. *arXiv preprint arXiv:2508.00161*, 2025.

## A. Additional Results

### A.1. Layer-wise Weight Change Magnitude

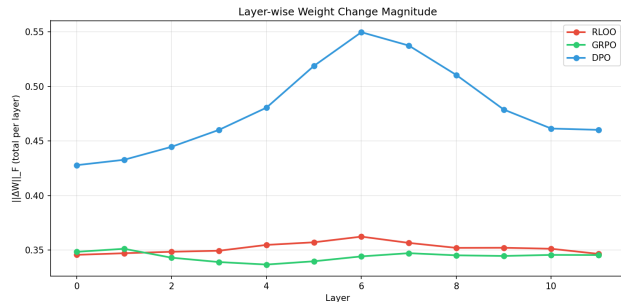


Figure 6. Total Frobenius norm of  $\Delta W$  per layer. DPO (blue) shows substantially larger magnitude changes, peaking at layers 5–7, while RLOO and GRPO show flat, lower magnitude.

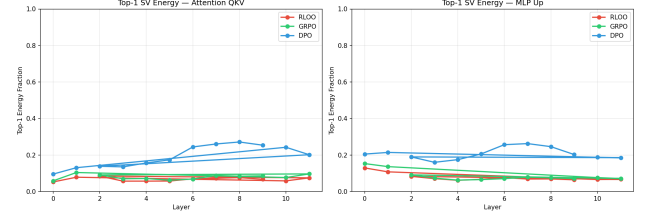


Figure 7. Fraction of  $\|\Delta W\|_F^2$  captured by the top-1 singular value, by layer. DPO (blue) consistently concentrates more energy in the top direction, especially in attention weights at layers 6–11.

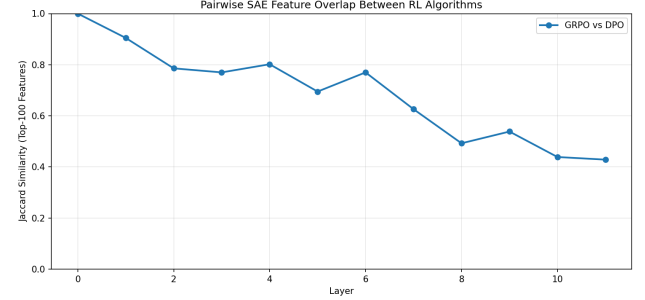


Figure 8. Pairwise SAE feature overlap between algorithms at 500 steps. RLOO–GRPO (blue) maintains higher overlap than either pair with DPO across all layers.

### A.2. Top-1 Energy by Layer

### A.3. Pairwise SAE Feature Overlap at 500 Steps

### A.4. Cross-Architecture: Qwen3-0.6B Effective Rank

### A.5. Toxicity Task: Effective Rank and SAE Overlap

## B. Experimental Details

### B.1. Training Hyperparameters

### B.2. SAE Details

We use the `gpt2-small-res-jb` SAE release from SAELens, which provides residual stream SAEs at all 12 layers of GPT-2 Small. Each SAE has 24,576 features with a target L0 of  $\sim 45$  active features per token.

### B.3. Compute

All experiments were run on the Harvard FASRC cluster using NVIDIA H100 80GB GPUs. Training each algorithm takes approximately 1–3 minutes on a single H100. Total compute: approximately 10 GPU-hours including all ablations, seeds, and analysis.



Figure 9. Effective rank of  $\Delta W$  by layer for Qwen3-0.6B (28 layers). DPO (blue) shows the same pattern as GPT-2: consistently lower rank that decreases sharply in deeper layers (dropping below 300 in layers 10–20), while RLOO and GRPO maintain higher rank.

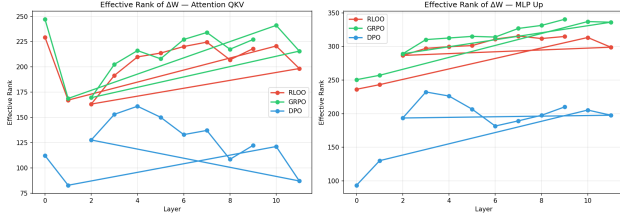


Figure 10. Effective rank on the toxicity task (GPT-2, 200 steps). The same rank ordering (online RL > DPO) holds with a different reward signal, confirming the pattern is task-independent.

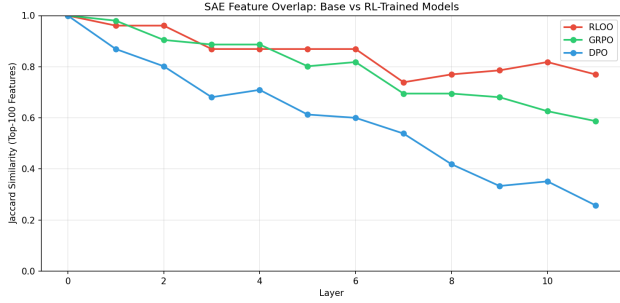


Figure 11. SAE feature overlap on the toxicity task. DPO (blue) diverges from base features faster than online RL, with dramatic drops in layers 8–11—the same cascade pattern as the sentiment task.

Table 5. Training configuration for all algorithms.

Parameter	RLOO / GRPO	DPO
Base model	GPT-2 Small (124M)	GPT-2 Small (124M)
Gradient steps	200	200
Learning rate	$1.41 \times 10^{-5}$ ( $1 \times 10^{-5}$ for RLOO)	$1.41 \times 10^{-5}$
Batch size	4 prompts $\times$ 4 gen.	16
Max new tokens	48	—
Max length	—	128
$\beta$ (KL / DPO)	0.04 (GRPO)	0.1
Precision	bf16	bf16
Optimizer	AdamW	AdamW
Max grad norm	0.5	0.5