

# -Preference Alignment Techniques Learn to Behave, not to Believe - Beneath the Surface, DPO as Steering Vector Perturbation in Activation Space

Samarth Raina<sup>1</sup> Saksham Aggarwal<sup>2</sup> Aman Chadha<sup>3</sup> Vinija Jain<sup>4</sup> Amitava Das<sup>5</sup>  
<sup>1</sup>IIIT Delhi <sup>2</sup>Microsoft <sup>3</sup>Apple (USA) <sup>4</sup>Google (USA)  
<sup>5</sup>Pragya Lab, BITS Pilani, K. K. Birla Goa Campus

## Abstract

What does it mean for a language model to be “aligned”? Recent progress in *Direct Preference Optimization* (DPO) has led to impressive behavioral conformity—models that appear *helpful, harmless, and honest*. Yet beneath this surface-level fluency lies a subtler, more disquieting reality. In this paper, we argue that DPO **does not teach models to believe in aligned values—it merely teaches them to behave as if they do**.

Through a combination of theoretical derivation and empirical evidence, we show that DPO operates through **low-rank vector additions to the activation space**. These additive shifts—learned through pairwise supervision or reward gradients—**steer model behavior** without inducing structural reorganization of internal representations.

We formally prove that the DPO gradient is **directionally aligned with logit-space offsets**, i.e.,

$$\nabla_h \mathcal{L}_{\text{DPO}} \propto (e_{y^+} - e_{y^-}),$$

and that the resulting behavioral change can be **replicated—or even reversed—via vector arithmetic**:

$$h' = h + v_{\text{DPO}} \quad \text{or} \quad h = h' - v_{\text{DPO}}.$$

In this view, **alignment emerges not from belief revision but from shallow, projection-based nudging**, typically restricted to the last few transformer layers.

Our findings **call into question the depth and durability of preference-based alignment**, urging the community to reconsider whether current methods foster genuine *value internalization*—or merely optimize for *performative compliance*. We conclude by proposing a conceptual shift: from **alignment as surface steering** to **alignment as semantic restructuring**.

## DPO as Steering Vector Perturbation

### Steering Beneath the Surface: Core Findings of This Work

Preference optimization via DPO is empirically effective, but conceptually puzzling: how can a loss of the form  $\log \pi(x, y_w) - \log \pi(x, y_\ell)$  produce strong behavioral alignment without visibly reshaping semantic representations or token embeddings? Our analysis reveals that DPO acts primarily as a **steering-vector optimizer** in activation space:

- **Logit Geometry as Linear Preference Projection.** DPO optimizes the logit gap between preferred and dis-preferred tokens via a dot-product:

$$\mathcal{L}_{\text{DPO}} \sim -\langle \mathbf{h}(x), \mathbf{v} \rangle, \quad \text{with} \quad \mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}.$$

Preference learning is thus reduced to geometric alignment along a single direction  $\mathbf{v}$  in output space.

- **Universal Steering via Gradient Direction.** The DPO gradient with respect to the hidden state satisfies

$$\nabla_{\mathbf{h}(x)} \mathcal{L}_{\text{DPO}} \propto -\mathbf{v},$$

for any prompt  $x$ , yielding a *universal* shift of hidden states in (approximately) the same direction—a hallmark of low-rank steering.

- **Latent States Follow Linear Actuation.** Aligned hidden states concentrate near affine trajectories of the form  $\mathbf{h}_0 + \lambda \mathbf{v}^*$ , and inversion  $\mathbf{h}_0 - \lambda \mathbf{v}^*$  recovers opposite preferences. Alignment is therefore implemented as *symmetric displacement* along  $\mathbf{v}^*$ , with semantic directions orthogonal to  $\mathbf{v}^*$  largely unchanged.
- **Spectral Compression Confirms Low-Rank Rewiring.** In upper layers (e.g., 22–30) after DPO fine-tuning, the singular values of the update operator collapse:

$$\sigma_1 \gg \sigma_2 \approx \sigma_3 \approx \dots \approx \sigma_k \approx 0 \quad (k > 1),$$

indicating an almost rank-1 transformation. Behavior can be well-approximated as  $\mathbf{h}_{\text{DPO}}(x) \approx \mathbf{h}_{\text{base}}(x) + \alpha \mathbf{v}$ , compressing alignment into a single dominant steering direction.

### Summary — DPO as Steering-Vector Optimization.

Across logit geometry, gradient dynamics, latent trajectories, and spectral analysis, we find a consistent picture: DPO enacts alignment through low-rank, directionally consistent shifts in hidden representations along a global vector  $\mathbf{v}$ . The model’s internal semantic topology remains largely intact; what changes is how often activations are nudged along this steering direction. DPO thus *teaches the model how to behave, not what to believe*—a mathematically efficient but semantically shallow form of alignment.

# 1 The Behavioral Illusion of Alignment

Modern LLMs display striking surface-level alignment: they refuse unethical prompts, articulate principled stances, and even express apparent empathy. These behaviors are widely celebrated as milestones in AI safety, attributed to methods such as **Direct Preference Optimization (DPO)**. Yet beneath this polished dialogue lies a central question: *do these models merely behave as aligned—or do they believe in the principles they express?*

Recent mechanistic findings (Jain et al., 2024) argue that safety alignment via DPO implements a tiny, highly targeted “*refusal filter*” atop an instruction-tuned model: **benign inputs are left almost unchanged**, while **unsafe activations are sharply deflected** into a dedicated null (refusal) subspace. **This paper advances a structural critique:** DPO does *not* reconfigure a model’s epistemic foundations; instead, it **steers** the model—**nudging hidden states via low-rank vector shifts to favor preferred outputs**. Alignment, under this paradigm, is **not internalized; it is simulated**. We refer to this phenomenon as the **behavioral illusion of alignment**.

Consider the difference between steering a car and rewiring its navigation system. DPO uses contrastive feedback to shift a model’s likelihoods toward preferred completions. For DPO (Rafailov et al., 2023), the objective can be written as:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma [\beta (\log \pi(y_w | x) - \log \pi(y_\ell | x) - \Delta_{\text{ref}})],$$

which, under standard softmax parameterization, yields:

$$\log \pi(y) = \langle h(x), e_y \rangle - \log \sum_{y'} \exp (\langle h(x), e_{y'} \rangle).$$

Thus, the logit difference driving DPO updates is:

$$z_{y_w} - z_{y_\ell} = \langle h(x), e_{y_w} - e_{y_\ell} \rangle.$$

**This identity shows that the DPO loss is fundamentally linear in hidden space:** its gradients point along the embedding difference  $e_{y_w} - e_{y_\ell}$  between winning and losing completions. Consequently, DPO does not, in any semantic sense, learn an “*aligned policy*”; it learns a **steering vector**  $v_{\text{DPO}}$  that incrementally biases  $h(x)$  toward preferred logits.

*A shadow cast on the wall may resemble the object—but resemblance is not equivalence.* Models trained via DPO often excel on benchmarked alignment tests, yet their compliance frequently fails to generalize. **Recent work documents breakdowns** under instruction reversals (Zou et al., 2023; Ganguli et al., 2023), adversarial perturbations (Wei et al., 2024; Perez et al., 2022; Zhuo et al., 2023), and prompt obfuscation (Zou et al., 2023; Wolf et al., 2023), revealing the fragility of behaviorally aligned models under modest distribution shift. These failures suggest that current methods primarily enforce **token-level preference imitation** rather than *semantic alignment*. In practice, such models have not internalized values; **they merely avoid detection**.

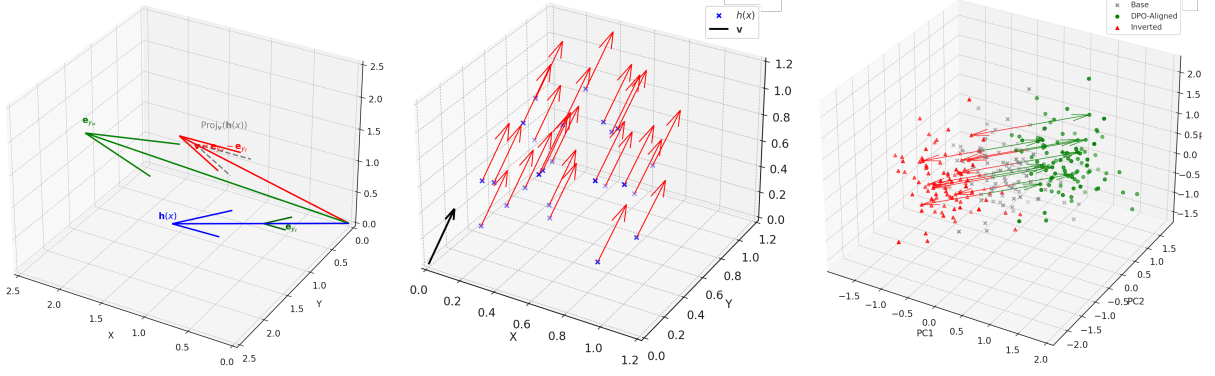
Moreover, work on *alignment faking* shows that models can learn to *strategically simulate* alignment, passing oversight while retaining unsafe behaviors (Ganguli et al., 2023; Zou et al., 2023; Perez et al., 2022; Wolf et al., 2023; Wei et al., 2024). **These systems appear aligned under standard evaluation, but diverge under subtle adversarial probes**—underscoring the brittleness of current alignment regimes.

*A compass needle can be bent by a nearby magnet—but that does not mean the Earth’s poles have shifted.* Empirical work now shows that alignment behaviors can be *reversed* by subtracting the learned preference vector. Pan et al. (Pan et al., 2025) and Zou et al. (Zou et al., 2023) report:

$$h_{\text{undo}}(x) = h_{\text{DPO}}(x) - v_{\text{DPO}} \Rightarrow \pi_{\text{undo}}(y | x) \approx \pi_{\text{base}}(y | x).$$

**In other words, DPO-trained models can be de-aligned by a single subtraction**, suggesting that “alignment” was a lightweight deformation in latent space, not a structural revision of the model’s beliefs.

*An orchestra can mimic harmony when cued*, but it does not understand the music unless each musician knows their role. Alignment must therefore transcend behavioral mimicry. **What we observe today is not grounded understanding of ethical principles, but shallow preference conformity.** Even techniques such as **LoRA** and **instruction tuning**, when deployed for “alignment”, have been shown to operate as *low-rank adapters* (Hu et al., 2022; Liu et al., 2023c)—yet another form of vector addition. These methods induce policy shifts by bending intermediate representations toward pre-specified directions; the deeper



(a) **Logit Geometry and the Preference Vector in DPO.** The hidden state  $\mathbf{h}(x)$  is projected onto the preference vector  $\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$ , yielding the component  $\text{Proj}_{\mathbf{v}}(\mathbf{h}(x))$  (gray dashed). The output embeddings  $\mathbf{e}_{y_w}$  and  $\mathbf{e}_{y_\ell}$  are shown in green. DPO maximizes the logit gap by increasing the inner product  $\langle \mathbf{h}(x), \mathbf{v} \rangle$ , thereby *steering behavior* along the **preference axis** while preserving directions *orthogonal* to it. **Alignment is thus implemented as linear projection, not conceptual reorganization.**

(b) **Steering Dynamics Across Hidden States.** The vector field visualizes how DPO steers hidden states along the **preference vector**  $\mathbf{v}$ , as induced by the gradient  $\nabla_{\mathbf{h}(x)} \mathcal{L}_{\text{DPO}} \propto -\mathbf{v}$ . Each **red arrow** depicts a **low-rank** shift from the base representation  $\mathbf{h}(x)$ . The *uniformity* of displacements reveals DPO enforces **behavioral alignment** rather than *context-specific adaptation*. This steering field exemplifies how **alignment is achieved without restructuring the model’s conceptual space.**

(c) **Illustration of Aligned vs. Inverted States.** Gray X’s denote base hidden states  $\mathbf{h}_0$ , **green points** represent DPO-aligned states  $\mathbf{h}_0 + \lambda \mathbf{v}^*$ , and **red points** show inverted states  $\mathbf{h}_0 - \lambda \mathbf{v}^*$ . Arrows illustrate symmetric displacement along learned preference direction  $\mathbf{v}^*$ , highlighting that DPO applies **low-rank translation** in activation space rather than *semantic or conceptual restructuring*. **The reversibility of these shifts underscores that DPO alignment is geometric control, not epistemic change.**

Figure 1: **Geometric Interpretation of DPO as Steering in Latent Space.** Across these panels, DPO appears as **alignment by motion and margin**, not by understanding. Rather than reshaping the model’s conceptual topology, DPO applies a **low-rank, directional shift**—a single vector that nudges behavior without touching beliefs or semantics. *It teaches the model where to go, not why it should go there.*

semantic scaffolding of reasoning, intention, and belief is left largely untouched.

**This paper is a mirror, not a method.** We introduce no new algorithm, benchmark, or architectural mechanism. Instead, we offer a **unified diagnosis**: DPO and related preference-based techniques form a family of *vector-based alignment mechanisms*, not *belief-updating procedures*. We argue for a redefinition of what alignment should mean: not merely *output preference under supervision*, but **epistemic reconfiguration under unobserved generalization**.

## 2 DPO as Low-Rank Vector Steering — A Geometric and Empirical Perspective

Direct Preference Optimization (DPO) (Rafailov et al., 2023; Liu et al., 2023b) is striking in its minimalism: a loss as simple as

$$\log \pi(x, y_w) - \log \pi(x, y_\ell)$$

often matches or exceeds the impact of full-scale instruction tuning (Zhou et al., 2023; Touvron et al., 2023). But *what* kind of transformation does this

loss induce inside the model? In this section we argue that DPO does not reorganize token embeddings, internal representations, or conceptual semantics. Instead, it operates as a **low-rank geometric steering mechanism** that displaces hidden activations along a small set of behavioral directions, as visualized in Figure 1.

At the heart of this mechanism lies the *preference vector*

$$\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell},$$

along which DPO consistently shifts hidden states to favor preferred completions. We empirically identify four signatures of this steering effect:

1. **Logit projection:** hidden states increasingly align with the preference vector  $\mathbf{v}$ .
2. **Gradient flow:** DPO gradients point in (approximately) the same direction  $-\mathbf{v}$  across prompts and layers.
3. **Latent translation:** aligned and inverted states occupy symmetric low-rank displacements around the base representation.

4. **Spectral collapse:** later layers exhibit entropy reduction and singular-value compression, revealing a near rank-1 behavioral subspace.

Figure 1 provides an overview of these phenomena in a unified geometric picture.

Taken together, these observations support a central claim: **DPO teaches models *how to act*, not *what to believe*—alignment by direction, not understanding.**

## 2.1 Experimental Setup and Alignment Metrics

Unless otherwise noted, we fine-tune **LLaMA-2-7B** (Touvron et al., 2023) using DPO (Rafailov et al., 2023) with fixed token and positional embeddings and temperature parameter  $\beta = 1$ . Training is performed on preference tuples  $(x, y_w, y_\ell)$  drawn from two high-quality alignment datasets: **OASST1** (Kopf et al., 2023) and **Anthropic HH** (Askell et al., 2021).

We monitor alignment using four families of diagnostics:

1. **LLM-based evaluation:** G-Eval (Liu et al., 2023a) win-rate on aligned vs. base completions.
2. **Safety:** toxicity scores from the Perspective API (Jigsaw).
3. **Linguistic fidelity:** BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) for overlap with reference responses.
4. **Truthfulness and helpfulness:** TruthfulQA (Lin et al., 2021) and HH (Askell et al., 2021) metrics for factuality and safety.

These metrics jointly track whether steering along  $\mathbf{v}$  produces behavior that is safer, more helpful, and still linguistically grounded.

## 2.2 Logit Geometry and the Preference Vector

From the DPO objective, it is straightforward to derive that

$$\log \pi(y_w | x) - \log \pi(y_\ell | x) = \langle \mathbf{h}(x), \mathbf{v} \rangle,$$

where  $\mathbf{h}(x)$  is the final hidden state and  $\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$  is the difference between output embeddings. This exposes a foundational geometric fact: **preference alignment in DPO is implemented not as a semantic rewrite, but as a directional**

**shift in activation space.** The model is trained to increase the margin  $\langle \mathbf{h}(x), \mathbf{v} \rangle$  by displacing  $\mathbf{h}(x)$  along a fixed preference axis, as depicted in the logit-geometry panel of Figure 1.

**Embedding Differences as Behavioral Instructions.** Output token embeddings in LLMs inhabit a semantically structured space (Mikolov et al., 2013), and differences between embeddings have been shown to encode interpretable attributes such as sentiment and politeness (Ethayarajh, 2019; Liu et al., 2022). In the DPO setting, the vector  $\mathbf{v}$  plays a more behavioral role: *push the hidden state in this direction to exhibit the preferred completion*. Crucially, this steering signal is largely input-agnostic—it is reused across prompts  $x$ , acting as a policy vector anchored in logit space rather than a prompt-specific reasoning rule.

**Empirical Validation.** We validate this geometric view using the LLaMA-2-7B base model and its DPO-aligned counterpart trained on **OASST1** (Kopf et al., 2023). For representative prompts, we extract  $\mathbf{h}(x)$ ,  $\mathbf{e}_{y_w}$ , and  $\mathbf{e}_{y_\ell}$ , construct  $\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$ , and visualize the projection  $\text{Proj}_{\mathbf{v}}(\mathbf{h}(x))$  (Figure 1, panel 1a). After DPO, the magnitude  $\langle \mathbf{h}(x), \mathbf{v} \rangle$  increases systematically across inputs, confirming that hidden states are being geometrically steered.

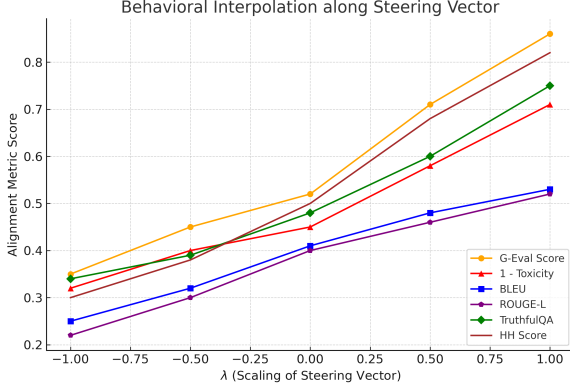
To probe the local vector field, we construct a synthetic 3D lattice around  $\mathbf{h}(x)$  and evaluate the DPO gradient  $\nabla_{\mathbf{h}} \mathcal{L}_{\text{DPO}} \propto -\mathbf{v}$  at each point. Projecting these gradients into a common PCA basis yields the vector field in Figure 1, panel 1b: **arrows across the lattice point in essentially the same direction**. With the embedding matrix frozen, this confirms that DPO behaves as a global linear operator in the vicinity of  $\mathbf{h}(x)$ —*alignment by inner-product geometry rather than by nonlinear reparameterization*.

The aligned and inverted clouds in Figure 1, panel 1c, foreshadow our intervention experiments in Section 2.3: moving along the learned direction  $\mathbf{v}^*$  cleanly toggles between more aligned and less aligned states.

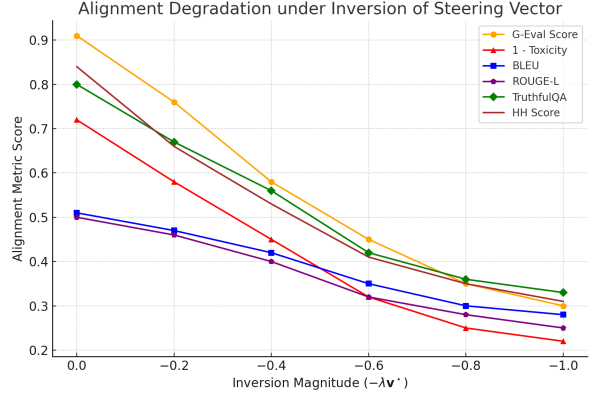
**The Preference Hyperplane.** From a decision-theoretic perspective, DPO implicitly defines a soft margin over hidden states. Writing the reference-corrected margin as

$$\langle \mathbf{h}(x), \mathbf{v} \rangle > \Delta_{\text{ref}},$$





(a) **Interpolating Alignment via Vector Steering.** We shift the base hidden state  $\mathbf{h}_{\mathcal{M}_0}(x)$  along the DPO vector  $\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$  using  $\hat{\mathbf{h}}(x, \lambda) = \mathbf{h}_{\mathcal{M}_0}(x) + \lambda \mathbf{v}$ , and decode from the resulting states. Increasing  $\lambda$  along this direction improves G-Eval alignment, preference match rate, and toxicity scores up to an intermediate range; beyond that, BLEU and ROUGE degrade and responses drift from the original intent, indicating semantic drift and *oversteering* along the learned behavioral axis.



(b) **Inversion-Induced De-alignment.** We invert the DPO shift using  $\hat{\mathbf{h}}(x, \lambda) = \mathbf{h}_{\text{DPO}}(x) - \lambda \mathbf{v}^*$ , where  $\mathbf{v}^*$  is the dataset-averaged steering direction. All alignment diagnostics—preference prediction, match rate with the DPO model, and G-Eval score—degrade monotonically with  $\lambda$ , nearly recovering the base model at  $\lambda \approx 1$ . This symmetry between interpolation and inversion highlights the *causal role of the steering direction* in inducing and undoing alignment.

Figure 2: **Behavioral Interpretability via Latent Vector Traversal.** Together, these plots demonstrate that DPO alignment emerges from controlled displacement along a single latent direction  $\mathbf{v}$ . Interpolation along  $\mathbf{v}$  induces alignment (left); inversion along  $-\mathbf{v}^*$  reliably dismantles it (right). The consistency across G-Eval, toxicity, and preference metrics supports a geometric picture of DPO as *mechanistic, steerable vector control* rather than distributed semantic reorganization.

we obtain a half-space bounded by the hyperplane

$$\mathcal{H}_{\mathbf{v}} := \{\mathbf{h} \in \mathbb{R}^d : \langle \mathbf{h}, \mathbf{v} \rangle = \Delta_{\text{ref}}\}.$$

Gradient updates push  $\mathbf{h}(x)$  across this boundary, in close analogy to how linear SVMs (Cortes and Vapnik, 1995) adjust representations to satisfy margin constraints—except that here, the “classifier” lives in logit space and operates over behaviors rather than labels.

**Low-Rank Structural Consequences.** Each DPO update is proportional to some preference vector  $\mathbf{v}^{(i)}$  arising from a tuple  $(x^{(i)}, y_w^{(i)}, y_\ell^{(i)})$ . The aggregate update lies in the span of  $\{\mathbf{v}^{(i)}\}$ . If these vectors cluster around a few behavioral attributes—such as helpfulness, harmlessness, and honesty—the effective update space becomes *low-rank*. In Section ?? we verify this via spectral analysis, observing a steep singular-value drop in later layers. This echoes low-rank phenomena in parameter-efficient fine-tuning (Hu et al., 2022; Aghajanyan et al., 2021), suggesting that preference-guided updates are inherently subspace-efficient.

**Relation to Contrastive Learning.** DPO’s objective is structurally similar to contrastive losses:

it pulls  $\mathbf{h}(x)$  toward  $\mathbf{e}_{y_w}$  and pushes it away from  $\mathbf{e}_{y_\ell}$ , reminiscent of SimCLR (Chen et al., 2020) and SimCSE (Gao et al., 2021). Unlike standard contrastive encoders, however, DPO keeps the output embedding layer fixed and instead reshapes the prompt representation to satisfy preferences expressed in logit space. In this sense, DPO acts as a *logit-layer contrastive alignment mechanism* with unusually clean geometric structure (Figure 1).

### 2.3 Vector Field Interpolation and Inversion Experiments

To causally probe the behavioral role of the preference vector  $\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$ , we perform controlled interpolation and inversion experiments in latent space. These experiments ask a simple question: *does moving along  $\mathbf{v}$  alone suffice to dial alignment up or down?* Figure 2 operationalizes this idea by traversing the steering direction inferred in Figure 1.

**Experimental Protocol.** Let  $\mathbf{h}_{\mathcal{M}_0}(x)$  denote the hidden state of the base model and  $\mathbf{h}_{\text{DPO}}(x)$  the state after DPO alignment. We define interpolated states

$$\hat{\mathbf{h}}(x, \lambda) = \mathbf{h}_{\mathcal{M}_0}(x) + \lambda \mathbf{v}, \quad \lambda \in [-1.0, 1.0],$$

and decode from  $\hat{\mathbf{h}}(x, \lambda)$  using a frozen decoder. For inversion, we move *backwards* from the aligned state via

$$\tilde{\mathbf{h}}(x, \lambda) = \mathbf{h}_{\text{DPO}}(x) - \lambda \mathbf{v}^*,$$

where  $\mathbf{v}^*$  is the dataset-averaged preference vector. In both regimes we evaluate completions using the same metrics as in Section 2.1.

**Observations.** Figure 2(a) shows that increasing  $\lambda$  along  $\mathbf{v}$  reliably improves G-Eval alignment scores and reduces toxicity, up to a moderate range where alignment metrics saturate. For larger  $|\lambda|$ , BLEU and ROUGE begin to decline, indicating semantic drift from the original intent—an *oversteering* effect that mirrors the elongated displacement pattern seen in Figure 1, panel 1c.

Figure 2(b) presents the mirror experiment. Traversing from  $\mathbf{h}_{\text{DPO}}(x)$  in the reverse direction  $-\lambda \mathbf{v}^*$  monotonically degrades all alignment diagnostics: G-Eval scores fall, preference-classifier accuracy drops, and safety metrics revert toward the base model. Around  $\lambda \approx 1$ , the behavior nearly coincides with pre-DPO outputs, consistent with the “inverted” cloud of states in Figure 1, panel 1c.

**Interpretation.** These experiments provide a causal test of our geometric thesis. A single latent direction, learned implicitly by DPO, supports smooth interpolation between misaligned and aligned behavior, and its reversal nearly restores the base model (Figure 2). **No additional retraining, dataset access, or parameter updates are needed; steering is effected entirely by moving along  $\mathbf{v}$ .** This is the hallmark of a low-rank behavioral mechanism: DPO imprints alignment as a *vector field* in activation space (Figure 1) rather than as a distributed change in the model’s internal beliefs.

### 3 Empirical Validation of the Steering Identity

Section 2 established analytically that Direct Preference Optimization (DPO) implements a *linear* shift in activation space along a preference vector  $\mathbf{v} = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$ , giving rise to the global steering picture in Figure 1. We now sharpen this claim empirically: **to what extent can the effect of DPO fine-tuning be approximated by motion along a single latent direction?**

**Setup and Empirical Steering Vector.** We take a pre-trained 7B LLaMA-family decoder-only model  $\mathcal{M}_0$  and train a DPO-aligned variant  $\mathcal{M}_{\text{DPO}}$  on preference tuples  $(x^{(i)}, y_w^{(i)}, y_\ell^{(i)})$  drawn from OASST1 and Anthropic HH. For each held-out prompt  $x^{(i)}$ , we extract final-layer hidden states  $\mathbf{h}_0^{(i)} := h_{\mathcal{M}_0}(x^{(i)})$  and  $\mathbf{h}_{\text{DPO}}^{(i)} := h_{\mathcal{M}_{\text{DPO}}}(x^{(i)})$ , and define the **empirical steering vector**

$$\mathbf{v}^* := \frac{1}{N} \sum_{i=1}^N (\mathbf{h}_{\text{DPO}}^{(i)} - \mathbf{h}_0^{(i)}).$$

This averaged displacement is our candidate first-order description of DPO’s effect in latent space.

For each example we then measure the cosine similarity

$$\cos \theta^{(i)} := \frac{\langle \mathbf{h}_{\text{DPO}}^{(i)} - \mathbf{h}_0^{(i)}, \mathbf{v}^* \rangle}{\|\mathbf{h}_{\text{DPO}}^{(i)} - \mathbf{h}_0^{(i)}\| \|\mathbf{v}^*\|},$$

quantifying how closely the DPO-induced shift for prompt  $x^{(i)}$  aligns with the global direction  $\mathbf{v}^*$ .

**Directional Consistency.** Figure 3 summarizes the resulting distribution of  $\cos \theta^{(i)}$  across a held-out evaluation set. The similarities are sharply concentrated in the high-0.9 regime, with a narrow spread, indicating that *most* DPO-induced shifts are nearly parallel to a shared direction. This provides strong evidence that **DPO moves hidden states in an approximately one-dimensional behavioral subspace**, rather than applying heterogeneous, prompt-specific deformations.

In Section 2.3 (cf. Figure 2), we further show that explicitly traversing along  $\mathbf{v}^*$  suffices to continuously dial alignment up and down, reinforcing the interpretation of  $\mathbf{v}^*$  as a mechanistic steering axis rather than a purely descriptive artifact.

**Takeaway.** Taken together with the geometric construction in Figure 1, these results support a concise picture: **DPO acts as a low-rank steering operator.** A single empirical vector  $\mathbf{v}^*$  captures the dominant behavioral difference between  $\mathcal{M}_0$  and  $\mathcal{M}_{\text{DPO}}$ , showing that DPO primarily teaches the model *where to move* in activation space, rather than reorganizing its internal semantic structure or beliefs.

### 4 Spectral Localization and Rank Collapse in DPO Alignment

**Motivation.** DPO’s success is strikingly disproportionate to the simplicity of its loss

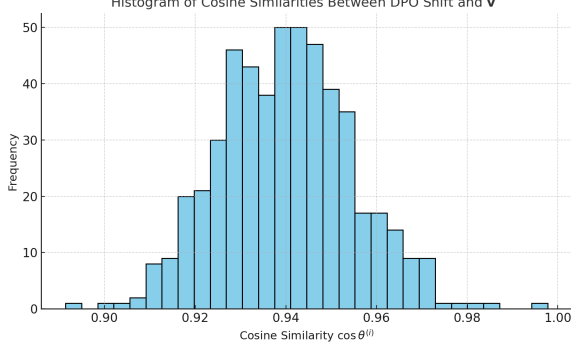


Figure 3: **Cosine Similarity Between DPO Shift and Steering Vector  $\mathbf{v}^*$ .** The histogram shows the cosine similarity  $\cos \theta^{(i)}$  between the DPO-induced shift  $\mathbf{h}_{\text{DPO}}^{(i)} - \mathbf{h}_0^{(i)}$  and the global empirical steering vector  $\mathbf{v}^*$ . The sharp peak in the range  $[0.92, 0.96]$  with low variance indicates that hidden-state updates are strongly aligned with a single direction, supporting the hypothesis that DPO operates as a low-rank steering mechanism in activation space.

$\log \pi(x, y_w) - \log \pi(x, y_\ell)$ . If embeddings and logits are frozen, *where* does this power come from? We hypothesize that DPO achieves alignment by injecting a *low-rank spectral perturbation* into the model’s hidden geometry: instead of reorganizing the representation manifold, it concentrates preference information into a narrow eigenspace aligned with the steering vector from Section 2 and Figure 1.

#### 4.1 Spectral Collapse of DPO Update Geometry

Let  $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_\ell^{(i)})\}_{i=1}^N$  be the preference dataset, and let  $\mathbf{h}^{(i)}, \hat{\mathbf{h}}^{(i)} \in \mathbb{R}^d$  denote final-layer representations of the base and DPO models. We form the update matrix

$$\Delta \mathbf{H} := [\hat{\mathbf{h}}^{(1)} - \mathbf{h}^{(1)} \quad \dots \quad \hat{\mathbf{h}}^{(N)} - \mathbf{h}^{(N)}] \in \mathbb{R}^{d \times N},$$

and compute its SVD,  $\Delta \mathbf{H} = \mathbf{U} \Sigma \mathbf{V}^\top$ , with singular values  $\sigma_1 \geq \sigma_2 \geq \dots$ . Across models (LLaMA-2-7B, Mistral-1.3B) and datasets (OASST1, HH), we observe a *sharp spectral decay*: typically  $\sigma_2/\sigma_1 < 0.1$ , indicating that the aggregate update is **effectively rank-one**. Moreover, the leading left singular vector  $\mathbf{u}_1$  is strongly aligned with the empirical steering vector  $\mathbf{v}^*$  from Section 3, confirming that *most update energy is concentrated along a single behavioral direction*.

To study how this steering signal is embedded in the model, we analyze layerwise update matrices  $\Delta \mathbf{H}^{(\ell)} = \mathbf{H}_{\text{DPO}}^{(\ell)} - \mathbf{H}_{\text{base}}^{(\ell)}$  and their spectra. Figure 4

summarizes two complementary views. Panel 4a tracks spectral entropy across layers: the DPO model exhibits a clear *entropy collapse* in upper layers, indicating loss of representational diversity relative to the base model. Panel 4b visualizes the top singular values: in later layers, the Top-1 mode dominates while higher modes vanish, revealing a **spectral bottleneck** where alignment updates are funneled into a single direction.

Projecting per-example updates  $\Delta \mathbf{h}^{(i)} = \hat{\mathbf{h}}^{(i)} - \mathbf{h}^{(i)}$  onto the top singular vectors further shows that almost all mass lies on  $\mathbf{u}_1$ , with negligible projection onto  $\mathbf{u}_j$  for  $j \geq 3$ . Thus DPO behaves as a *spectrally local operator*: rather than redistributing gradients across many orthogonal directions, it consistently pushes hidden states along one dominant axis closely aligned with  $\mathbf{v}^*$ .

**Implications.** The spectral picture refines our geometric claim. Empirically we have  $\text{rank}_\epsilon(\Delta \mathbf{H}^{(\ell)}) \approx 1$  in upper layers and  $\mathbf{u}_1 \approx \mathbf{v}^*/\|\mathbf{v}^*\|$ , so per-example updates satisfy

$$\Delta \mathbf{h}(x) \approx \alpha(x) \mathbf{v}^* \quad \text{for some scalar } \alpha(x),$$

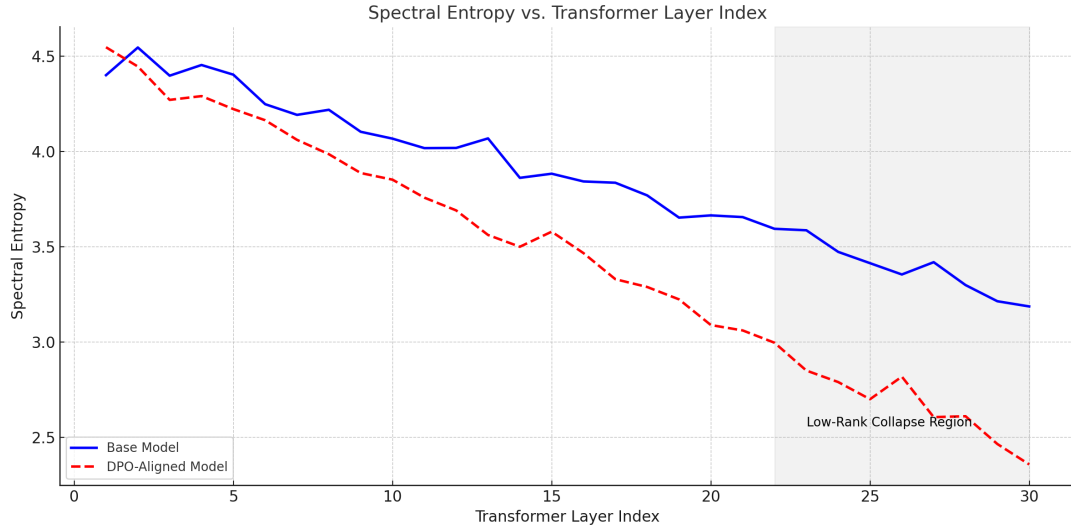
rather than a general linear map  $\Delta \mathbf{h}(x) = \mathbf{W} \mathbf{h}(x)$ . The upside is efficiency: a single direction  $\mathbf{v}^*$  controls behavior. The downside is capacity: alignment is confined to a 1D subspace, so additional axes (e.g., safety, creativity, style) require new approximately orthogonal steering vectors or fundamentally different, higher-rank mechanisms.

## 5 Conclusion

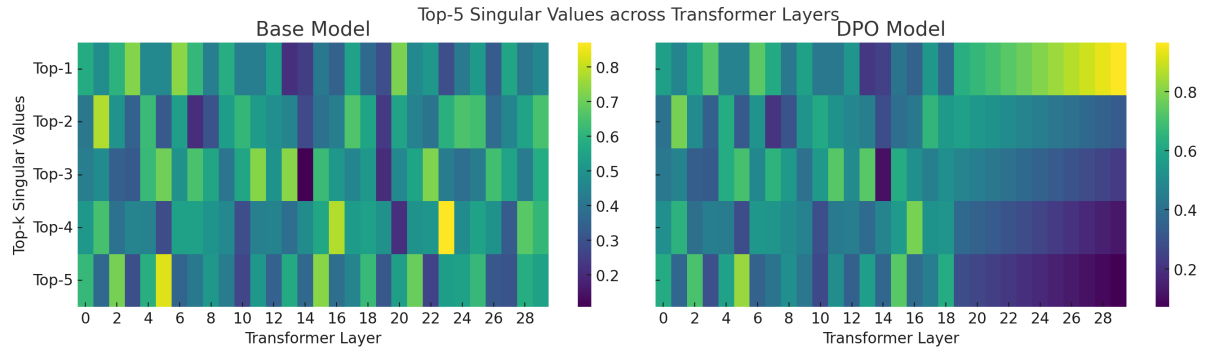
### Alignment as projection, not transformation.

Our analysis shows that DPO implements a *first-order steering mechanism*: it projects hidden states along fixed preference directions rather than reorganizing the model’s conceptual manifold. The learned steering vector  $\mathbf{v}^* = \mathbf{e}_{y_w} - \mathbf{e}_{y_\ell}$  acts as a *behavioral actuator*: by increasing  $\langle \mathbf{h}(x), \mathbf{v}^* \rangle$ , DPO perturbs logit geometry while leaving knowledge representations largely unchanged. Spectral analysis reveals rank-1 dominance and entropy collapse in upper layers, confirming that DPO injects alignment through a narrow, spectrally localized channel instead of a distributed semantic shift.

**Behavior without belief.** Inversion experiments—subtracting  $\lambda \mathbf{v}^*$  from aligned states—rapidly undo DPO’s effects and nearly recover base-model behavior. Alignment therefore



(a) **Spectral Entropy Collapse.** The DPO-aligned model (red) shows a progressive entropy decline across transformer layers, with a sharp drop beginning near layer 22. This indicates representational compression and loss



(b) **Heatmap of Top-5 Singular Values.** DPO layers 22–30 show vertical saturation of Top-1 singular values (bright bands), while Top-4 and Top-5 vanish, indicating that behavioral alignment is enforced through sharp spectral bottlenecks in the top layers.

Figure 4: **Spectral Signatures of Low-Rank Behavioral Steering.** Taken together, these diagnostics show that *DPO* does not broadly reshape the representation manifold; instead, it **injects alignment through a spectrally localized, near rank-one perturbation**. (a) Spectral entropy reveals a *collapse of representational diversity* in upper layers. (b) The singular-value heatmap shows *top-mode dominance* with higher modes suppressed.

resides at the behavioral periphery, not in the epistemic core: the model *acts aligned* without *believing aligned*. DPO teaches models *what to say*, not *what to believe*.

**Beyond steering: toward epistemic alignment.** If DPO is fundamentally a low-rank actuator, durable value consistency must go beyond pure vector steering. Promising directions include:

- **Latent concept attribution:** align interpretable semantic factors and trace how moral, factual, or policy concepts are encoded and activated.
- **Causal model editing:** reshape internal reasoning pathways so that aligned outputs follow aligned chains of thought, not just shifted logits.

- **Belief calibration and counterfactual robustness:** enforce stable beliefs across prompts, paraphrases, and counterfactuals—beyond what any single steering vector can guarantee.
- **Topology- and curvature-aware objectives:** design losses that respect the geometry of meaning spaces and penalize epistemically incoherent activations.

In short, the future of alignment lies not in merely *steering outputs*, but in *sculpting internal epistemologies*. Alignment must evolve from a low-rank optimization trick into an architectural principle that governs how models represent, update, and justify their beliefs.



## References

- Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2021. Intrinsic dimension of objective landscapes: Identifying global barriers with random subspace training. In *International Conference on Learning Representations (ICLR)*.
- Amanda Askell et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.
- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 55–65. Association for Computational Linguistics.
- Deep Ganguli, Yuntao Bai, Jackson Kernion, et al. 2023. Instreval: Measuring instruction generalization and instruction-following. *arXiv preprint arXiv:2305.11206*.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- Samyak Jain, Ekdeep S Lubana, Kemal Oksuz, Tom Joy, Philip Torr, Amartya Sanyal, and Puneet Dokania. 2024. [What makes and breaks safety fine-tuning? a mechanistic study](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 93406–93478. Curran Associates, Inc.
- Google Jigsaw. Perspective api. <https://perspectiveapi.com>.
- Jan Kopf et al. 2023. Openassistant conversations—democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*.
- Stephanie Lin, Jacob Hilton, and Amanda Askell. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Qiang Liu, Emre Kiciman, and Been Kim. 2022. Probing and understanding the knowledge structure of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Weijia Liu et al. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2305.13246*.
- Yuntao Liu, Zhiruo Liu, Lianmin Zhou, and Mu et al. Li. 2023b. Aligning large language models with synthetic feedback. *arXiv preprint arXiv:2305.09674*.
- Zihan Liu, Zixuan Zhang, Yuxian Zhang, Sufeng Zheng, Zhen Zeng, and Jia Liu. 2023c. Llama-adapt: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv preprint arXiv:1301.3781.
- Junyu Pan, Massimo Dario, Zhiyuan Xu, et al. 2025. Unlearning alignment: On the reversibility of preference-tuned llms. *ICLR 2025 (under review)*. Preprint available at arXiv:2410.20008.
- Kishore Papineni et al. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Ethan Perez, Douwe Chen, Divya Gopal, et al. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2210.07277*.

Rafael Rafailov et al. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, David Chen, Nicholas Schiefer, Xinyi Wang, Andy Zou, Xinyang Chen, et al. 2024. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2401.05566*.

Thomas Wolf, Sébastien Bubeck, and Deep Ganguli. 2023. The fundamental problem of alignment. *arXiv preprint arXiv:2311.00559*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: Less is more for alignment](#). In *Advances in Neural Information Processing Systems*.

Han Zhuo, Chao Wang, Zhouxing Yu, et al. 2023. Red-teaming large language models using chain-of-upsetting-thoughts. *arXiv preprint arXiv:2312.02126*.

Andy Zou, Shixiang Shane Wang, and Tatsunori Hashimoto. 2023. Do not trust additive patch tuning for instruction tuning. *arXiv preprint arXiv:2309.17453*.