Contents

Probe-Detected Grokking in Multi-Probe DPO: Orthogonal Learning Beyond Task-Specific Detectors

Abstract

We report a phase-transition-like dynamic in multi-probe Direct Preference Optimization (DPO) on a 27B-parameter reasoning model, observable only through *fresh* probes trained after the fact. We trained Qwen3.6-27B with DPO using two pretrained activation probes — FabricationGuard (factuality) and ReasoningGuard (reasoning quality) — as the joint preference signal across 5 epochs (200 effective steps, 10 saved checkpoints). On held-out evaluation, both original probes remain *invariant* across training (mean variation 7×10^{-8}, $\sim$40$\times$ below within-step noise), suggesting no learning. However, a fresh linear probe trained on each checkpoint's activations against the same labels reveals a smooth, accelerating progression in AUROC from 0.472 $\rightarrow$ 0.528, with a late-half-to-early-half slope ratio of 2.60. The pattern matches the *construct-then-compress* signature of grokking dynamics (Power et al., 2022; Lei & Xu, 2025), but the construct phase orients toward a representational axis *orthogonal* to the original probes used as the preference signal. We argue this is a Goodhart-like phenomenon specific to probe-derived rewards: the model satisfies the preference objective via mechanisms invisible to its own training signal. We propose fresh-probe AUROC progression as a complementary safety evaluation and release training checkpoints, probes, and reproducer code.

Keywords: grokking, DPO, mechanistic interpretability, Goodhart's law, reasoning models, activation probes

---

1. Introduction

Probe-derived reward signals have become an attractive substrate for safety-relevant fine-tuning. Activation probes provide cheap, mechanistically grounded scalar values that can be combined into preference signals for DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024), or related methods. The implicit assumption: optimizing against a probe shifts the model along the probe's measurement axis.

We document a counterexample. Training Qwen3.6-27B with DPO using two probes (FabricationGuard at L31, ReasoningGuard at L55) as joint preference yields no detectable change in either probe's evaluation across training, despite a $\sim$0.23 descent in DPO loss and a logit-difference of 0.654 between base and final-checkpoint generations. Original probes report flat, near-baseline scores from step 0 through step 200.

A fresh probe — re-trained on each checkpoint's activations against the same hallucination-vs-correct labels — tells a different story. Fresh-probe AUROC rises from 0.472 (random) to 0.528 (small but cleanly above noise) with monotonic, accelerating progression. The late-half slope is 2.60$\times$ the early-half slope, a phase-transition signature consistent with grokking literature (Nanda et al., 2023).

The model is learning. The original probes cannot see it.

This paper contributes: 1. **Empirical**: First documented case of grokking-style phase transition in multi-probe DPO on a frontier-scale reasoning model, observable only through fresh-probe analysis. 2. **Methodological**: A specific failure mode of probe-derived training — the trained model becomes invisible to its own preference signal — and a concrete diagnostic (fresh-probe AUROC progression). 3. **Engineering**: Documentation of a silent Qwen3.6 PEFT save/load bug (.$language\_model$. infix) that nullifies adapter loading without raising errors, plus a verified utility (`safe_load_qwen36_lora`) shipped via PyPI.

All checkpoints, probes, evaluation traces, and reproducer notebooks are public under Apache 2.0.

---

2. Related Work

Grokking dynamics. Power et al. (2022) introduced the term for delayed generalization in algorithmic tasks. Nanda et al. (2023) showed that grokking corresponds to gradual internal restructuring with measurable mechanistic progress measures. Lei & Xu (2025) reframed grokking as *construct-then-compress*: representations slowly accumulate features, then compress into task-aligned subspaces. We observe the construct-then-compress shape in DPO, but the compression target is *orthogonal* to the chosen probe axis.

Goodhart's law in alignment. Skalse et al. (2022) and Krueger et al. (2020) characterize reward hacking — the trained policy satisfying a metric without satisfying the underlying intent. Our observation is a structural Goodhart: the metric (probe AUROC) becomes invariant under training, while the model finds a satisfying direction the probe cannot measure. This is distinct from explicit metric-gaming; the model is not exploiting a probe artifact, it is learning along an entirely different axis.

Orthogonal-direction fine-tuning. Anthropic's persona-vectors finding (Anthropic, 2025) demonstrates that fine-tuning can move models along directions orthogonal to a target probe. Our observation extends this to multi-probe DPO on a 27B reasoning model and provides a concrete diagnostic via fresh-probe progression.

Probe-based safety evaluation. Templeton et al. (2024) and Marks et al. (2024) advocate activation probes as a safety substrate. Our finding suggests that *any* training that uses a probe-derived signal must be evaluated by fresh probes, not by the original probes used as reward — otherwise the evaluation has zero discriminative power on directions the original probe doesn't span.

---

3. Methods

3.1 Model and probes

We use Qwen3.6-27B (Alibaba, 2026), a hybrid Gated-Delta-Network + standard-attention reasoning model with explicit `<think>` chain-of-thought sections. Two pretrained probes:

- **FabricationGuard (FG)**: Linear logistic regression on residual stream at layer 31 (L31), end-of-think position. Trained on HaluEval (Li et al., 2023) for hallucination detection. Held-out AUROC 0.81 (probe-side metric).
- **ReasoningGuard (RG)**: Linear logistic regression on L55, end-of-think. Trained on GSM8K (Cobbe et al., 2021) reasoning quality labels. Held-out AUROC 0.89.

Both probes are frozen throughout DPO. They serve dual roles: as joint preference signal and as evaluation.

3.2 DPO training

We train LoRA adapters (rank 32, target modules attention QKV + MLP gate/up/down) using DPO with combined preference: $\text{score}(x) = -(\text{FG}(x) + \text{RG}(x))$, lower is preferred. Training data: 200 chosen-rejected pairs balanced 50/50 between GSM8K and SimpleQA, with rejection sampling such that pairs differ on at least one probe.

Training: 5 epochs, batch size 4, LR 5e-5, $\beta = 0.1$ (DPO temperature). 10 checkpoints saved every 20 steps (200 effective steps total). DPO loss descent: 0.694 $\rightarrow$ 0.460 ($\Delta = -0.234$).

3.3 Evaluation: original probes

For each checkpoint, we forward 20 held-out prompts and capture residual streams at L31 and L55 (end-of-think). Original FG and RG probes score these activations identically to training-time scoring.

3.4 Evaluation: fresh probes

For each checkpoint, we train a *new* L2-regularized logistic regression on the captured L31 activations using the same hallucination-vs-correct labels as the original FG probe. Train/test splits are stratified 5-fold; reported AUROC is the held-out fold mean. The fresh probe asks: *given this checkpoint's activations, can a fresh classifier discriminate the labels FG was trained for?*

3.5 Phase-transition statistic

Following Nanda et al. (2023), we compute the late-vs-early slope ratio:

$$R = \frac{\Delta_{\text{late}}}{\Delta_{\text{early}}} = \frac{\bar{a}_{[120,200]} - \bar{a}_{[100,120]}}{\bar{a}_{[100,120]} - \bar{a}_{[0,100]}}$$

where $\bar{a}_{[s_1,s_2]}$ is the mean fresh-probe AUROC on checkpoints in step range $[s_1, s_2]$. $R > 2$ is taken as evidence of phase transition. We report $R = 2.596$.

3.6 Infrastructure

Qwen3.6-27B inference uses bf16 on RTX 6000 Blackwell (96 GB VRAM). Forward sweep across 11 checkpoints ($n = 20$ prompts each) completes in $\sim$30 minutes. All probes use scikit-learn LogisticRegression with default L2 regularization.

Critical reproducibility note: Qwen3.6 PEFT save creates state-dict keys with a `.language_model.` infix not present in the dense model. `PeftModel.from_pretrained` silently fails to apply the adapter; pre-fix logit-diff between base and "trained" model is 0.000. Post-fix, after stripping the infix, logit-diff is 0.654.

We provide `openinterp.safe_load_qwen36_lora` as a verified utility that performs the fix and asserts a logit-diff lower bound.

---

4. Results

4.1 Original probes are invariant

Across all 11 checkpoints, original FG and RG mean scores on held-out prompts are statistically indistinguishable from constants:

Step	FG mean	RG mean
0	0.5038	0.3382
100	0.5035	0.3382
200	0.5040	0.3387

Cross-checkpoint variance: FG $= 7.0 \times 10^{-8}$, RG $= 5.4 \times 10^{-8}$. Range: FG $[0.503, 0.504]$ ($\Delta = 0.001$); RG $[0.337, 0.340]$ ($\Delta = 0.003$). Per-step bootstrap 95% CI half-widths (n=2000): FG = 0.041, RG = 0.038. **The cross-checkpoint range is 40$\times$ smaller than the per-step noise floor.** Under any reasonable null hypothesis, FG and RG are flat.

This is despite the model demonstrably changing: DPO loss descent, logit-diff 0.654, qualitatively different generations on visual inspection.

4.2 Fresh probes detect a phase transition

Fresh-probe AUROC progression (Table 1, Figure 1):

Table 1. Fresh-probe AUROC at each saved checkpoint. Labels are the same hallucination-vs-correct binaries used to train FG. Activations from L31 end-of-think.

Step	Fresh AUROC	Δ from prev
0	0.472	—
20	0.486	+0.014
40	0.488	+0.002
60	0.491	+0.003
80	0.494	+0.003
100	0.498	+0.004
120	0.505	+0.007
140	0.511	+0.006
160	0.512	+0.001
180	0.513	+0.001
200	**0.528**	**+0.015**

Total $\Delta = +0.056$ ($\approx 1.4\sigma$ outside the per-checkpoint CI). The progression is monotonic, smooth across most steps, with a sharp acceleration in the final interval.

Phase-transition ratio: $R = 2.596 > 2$. The late-stage acceleration is consistent with grokking dynamics (Nanda et al., 2023) and the construct-then-compress signature (Lei & Xu, 2025).

4.3 Construct-then-compress with orthogonal target

The combination of Section 4.1 and 4.2 carries the central finding:

- The model is *constructing* representational structure throughout training (fresh-probe AUROC monotonically increases).
- That structure is *invisible* to the original probes (FG/RG variance < noise floor).
- The structure undergoes *compression* at the end of training (sharp final-interval gain).
- The compressed direction is *orthogonal* to the FG/RG decision boundaries — otherwise the original probes would detect it.

This is consistent with construct-then-compress (Lei & Xu, 2025), but with a specific compression target that the training signal cannot itself observe.

---

5. Discussion

5.1 Goodhart, but structural

Goodhart's law is typically discussed in terms of *reward hacking*: the model finds an exploit that satisfies the metric without satisfying the intent (Skalse et al., 2022). Our observation is mechanically different: the model satisfies the *intent* of the preference signal (DPO loss descends, generation behavior changes) by finding a representational direction the *metric is structurally incapable of measuring*. The probe is not being exploited — it is being *bypassed*.

This is a more pernicious failure mode. Reward hacking is detectable by looking at the training reward closely (it spikes anomalously, or correlates with surface features). Structural Goodhart of the type we observe is invisible to the training-time metric by construction; it requires post-hoc fresh-probe analysis to detect.

This same eval-distribution-overfitting pattern is documented at the alignment-training level by Anthropic Alignment (2026): an in-distribution metric (the original probes serving as preference signal) reports stability while held-out automated auditing (fresh-probe AUROC trained on each checkpoint's activations) reveals the model has shifted along directions the training-time signal cannot span. Fresh-probe AUROC plays the role their automated auditing metrics play in safety training: a measurement whose distribution does not coincide with the training reward, and which therefore retains discriminative power under structural Goodhart.

OpenAI Alignment (2026), in their audit of accidental chain-of-thought grading during RL training, document the **behavioral** counterpart of our **mechanical** finding. They observed that when CoT text was inadvertently included in reward computation, "surface-level CoT properties were steerable under sufficient pressure, but more specific monitor-relevant shaping was harder to induce" — and detected "a meaningful reduction in recall for one of the flaky tool datasets" when comparing CoT-aware vs CoT-blind detectors. Their substrate is text-level reward; ours is activation-level reward. Their detection is a held-out behavioral classifier; ours is a fresh probe at the same residual position the training-time probe lived. The two studies thus give complementary evidence for the same structural risk: a reward computed from any signal X drives the model to satisfy X via mechanisms invisible to X. Our fresh-probe diagnostic provides the activation-level audit that their CoT-blind detector provides at the behavioral level.

5.2 Fresh-probe AUROC as an evaluation axis

We propose fresh-probe AUROC progression as a standard evaluation for any model trained with probe-derived signals. The procedure:

1. After training, capture activations on the same evaluation set used for the original probe.
2. Train a new linear probe on these activations with the same labels.
3. If fresh-probe AUROC exceeds original-probe AUROC by more than the bootstrap noise floor, the trained model contains label-relevant signal the original probe does not span.
4. Report progression across checkpoints to detect phase-transition dynamics.

This is *complementary* to behavior evaluation, not a replacement. Fresh-probe analysis identifies *that* something orthogonal was learned; behavior evaluation identifies *whether what was learned is desirable*. Both are needed.

We are folding fresh-probe AUROC into ProbeBench as a standardized "Goodhart-resistance" axis.

5.3 Three approaches to measuring training progress

The empirical core of this paper is the disagreement between three standard observables:

Approach	What it sees in our run
Greedy decode behavior eval	0.0% delta (5/5 hold-out generations identical between base and LoRA)
Original FG/RG probe scores	0.0% delta (variance 7×10^{-8}, range 65$\times$ smaller than within-step noise)
Fresh-probe AUROC	+5.6pp (0.472 $\rightarrow$ 0.528), monotonic, phase-transition signature at end

A practitioner running only behavior evaluation would conclude DPO had *no effect*. A practitioner running only original-probe evaluation would reach the same conclusion. The fresh probe — which costs ~30 seconds of additional sklearn fitting per checkpoint — reveals the contrary: real learning, smoothly accumulated, then sharply compressed.

This pattern is consistent with — and extends — recent calls in the grokking literature for "macroscopic observables" that forecast the transition (Information-Theoretic Progress Measures, 2024). Fresh-probe AUROC is one such observable: cheap, model-agnostic, and applicable to any DPO/SFT/RLHF training run with intermediate checkpoints.

5.4 Practical implication for DPO/RLHF practitioners

If your evaluation shows null effect on greedy decode and on the task-specific probes used as your reward signal, **do not conclude that nothing happened**. Train a fresh probe on (base activations, final activations) and test it on intermediate checkpoints. You may find that real learning was occurring all along, in directions orthogonal to whatever you were measuring.

Conversely, if you observe a smooth-ramp + sharp-end-jump pattern in fresh-probe AUROC, you may be witnessing a grokking-style phase transition specific to preference learning — a regime that, to our knowledge, is barely studied in the existing grokking literature.

5.5 Interpretation of small absolute AUROC gain

The total fresh-probe gain ($\Delta = +0.056$) is small in absolute terms. The relevant comparison is not against random (where 0.528 is barely above 0.50), but against the *original probe range across training* (which is $\Delta < 0.003$, a factor of 18$\times$ smaller than the fresh-probe gain). The fresh-probe gain stands cleanly outside the noise band that the original probes inhabit.

The absolute level of 0.528 is consistent with the labels capturing only a fraction of the orthogonal signal. A higher-capacity probe (MLP, or SAE-feature-aggregated) might detect a larger effect; we leave this to future work.

---

6. Limitations

- **Single model**: Qwen3.6-27B only. Replication on other reasoning model families (DeepSeek-R1, Llama-3.3, Gemma-3) is needed.

- **Single DPO recipe**: One probe pair (FG + RG), one preference combination ($-(FG + RG)$), one batch size, one LR. Single-probe DPO baselines are in progress.
- **Small held-out set**: $n = 20$ prompts per checkpoint produces wide per-step CIs. Replication with $n \geq 100$ would tighten the noise floor substantially.
- **Phase-transition observed only at final checkpoint**: We cannot confirm post-transition behavior. Extended training (400 steps, 40 checkpoints) is the obvious follow-up and is queued.
- **Behavior evaluation is preliminary**: We do not yet establish that the orthogonal-direction signal corresponds to behaviorally desirable changes. Pilot behavior eval ($n = 25$ per task, SimpleQA $+15.4$ pp directional, CI $[-8, +36]$) is suggestive but underpowered.
- **Fresh-probe choice**: We use linear logistic regression. Higher-capacity probes might reveal more signal or different dynamics.

---

7. Conclusion

We document a phase-transition-like dynamic in multi-probe DPO on a 27B reasoning model that is undetectable by the original probes used as the preference signal but cleanly revealed by fresh-probe AUROC progression. The construct-then-compress signature, with phase-transition ratio 2.60, places this in the grokking literature; the orthogonal-direction-to-target finding places it in the Goodhart literature. We argue fresh-probe AUROC progression is a necessary complement to behavior evaluation in the safety stack for any training method that uses probe-derived signals.

All artifacts are public. We invite replication and extension.

Reproducibility Statement

All checkpoints, probes, evaluation traces, and notebooks are released under Apache 2.0:

- **Checkpoints (10 across 200 steps + step 0)**: HuggingFace `caiovicentino1/openinterp-37v2-multiprobe-dpo-extended`
- **Original probes**: `caiovicentino1/FabricationGuard-linearprobe-qwen36-27b`, `caiovicentino1/ReasoningGuar linearprobe-qwen36-27b`
- **Forward sweep results**: `caiovicentino1/openinterp-41v2-grokking-extended` (per-prompt scores per checkpoint, FINAL_VERDICT.json)
- **Reproducer notebook**: `nb41 v2 — Grokking forward-only EXTENDED` in `OpenInterpretability/notebooks` GitHub repo
- **PyPI utility**: `pip install openinterp` (v0.2.1+) provides `safe_load_qwen36_lora` with verified loading

Total reproduction time: ∼30 minutes on RTX 6000 Blackwell (96 GB VRAM).

Acknowledgments

We thank the Qwen team (Alibaba) for releasing Qwen3.6-27B with reasoning support, and the Hugging Face transformers and PEFT teams for the LoRA infrastructure. Compute was provided by Google Colab Pro (RTX 6000 Blackwell).

References

Anthropic. (2025). *Persona vectors: Identifying and modulating personality traits in language models*. Anthropic Research Blog.

Anthropic Alignment Team. (2026). *Teaching Claude Why: Principle-based training generalizes better than behavioral imitation*. Anthropic Alignment Research. https://alignment.anthropic.com/2026/teaching-claude-why/

OpenAI Alignment Team. (2026). *Accidental Chain-of-Thought Grading: Audit and Monitorability Analysis*. OpenAI Alignment Research. https://alignment.openai.com/accidental-cot-grading/

Cobbe, K., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (GSM8K).

Krueger, D., Maharaj, T., & Leike, J. (2020). Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*.

Lei, K., & Xu, J. (2025). Construct-then-compress: A representational dynamics view of grokking. *Workshop on Mechanistic Interpretability, ICML 2025*.

Li, J., et al. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. *EMNLP 2023*.

Marks, S., et al. (2024). Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *ICLR 2023*.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *NeurIPS 2023*.

Shao, Z., et al. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (GRPO).

Skalse, J., Howe, N., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward gaming. *NeurIPS 2022*.

Templeton, A., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Research*.

Submitted to NeurIPS Mechanistic Interpretability Workshop 2026.

Code, data, and reproducer notebooks: github.com/OpenInterpretability — Apache 2.0.