

Embedding Norm Heterogeneity Drives LoRA Fine-Tuning Brittleness in Overtrained Language Models

Anonymous
Anonymous Institution

March 2026

Abstract

Extended pre-training makes language models harder to fine-tune—a phenomenon termed “catastrophic overtraining” (Springer et al., ICML 2025). We investigate the geometric mechanisms behind this brittleness by tracking the LM head geometry across Pythia training checkpoints. Starting from the hypothesis that ETF (Equiangular Tight Frame) crystallization drives fine-tuning difficulty, we instead discover that **embedding norm heterogeneity** is the causal geometric predictor. Our key findings: (1) ETF overlap geometry does *not* predict LoRA fine-tuning difficulty ($r = 0.14$); (2) the coefficient of variation of LM head embedding norms strongly predicts deconfounded fine-tunability ($r = -0.84$ at 410M, $r = -0.99$ at 1.4B); (3) equalizing norms before LoRA increases relative fine-tunability by up to 79% for late-training checkpoints, with the effect scaling monotonically with norm CV; (4) the effect is rank-independent—increasing LoRA rank does not mitigate it. We show that norms encode semantic specificity rather than frequency, and that the resulting impedance mismatch between high- and low-norm tokens creates a fundamental barrier for uniform low-rank adaptation.

1 Introduction

Two recent results in the theory of large language models connect naturally but have not been linked mechanistically.

Superposition and ETF structure. Liu et al. [1] (NeurIPS 2025 Best Paper Runner-Up) demonstrated that LLM heads operate in “strong superposition,” with feature vectors forming Equiangular Tight Frames (ETFs) where pairwise overlaps converge to $1/\sqrt{d}$. This geometric structure makes the scaling law $L \propto 1/m$ a geometric inevitability of compressing sparse concepts into dense spaces.

Catastrophic overtraining. Springer et al. [2] (ICML 2025) showed that extended pre-training can make models *harder* to fine-tune, terming this “catastrophic overtraining.” For example, OLMo-1B pre-trained on 3T tokens performs over 2% worse after instruction-tuning than its 2.3T token counterpart. They attribute this to increased “broad sensitivity” of pre-trained parameters.

The gap. Neither work identifies the *geometric mechanism* connecting pre-training dynamics to fine-tuning difficulty. If the LM head crystallizes into a tighter ETF during training, does this geometric rigidity explain why LoRA fine-tuning becomes harder?

Our contribution. We track LM head geometry across Pythia [3] training checkpoints and discover that the answer is *no*—ETF overlap geometry does not predict fine-tuning difficulty. Instead, we identify **embedding norm heterogeneity** as the causal geometric mechanism. Specifically:

1. **Null result:** ETF overlap metrics (mean, variance of pairwise overlaps) show no correlation with LoRA fine-tunability ($r \approx 0.2$).
2. **Predictive:** The coefficient of variation (CV) of LM head row norms predicts deconfounded fine-tunability with $r = -0.99$ at the 1.4B scale.
3. **Causal:** Equalizing norms before LoRA fine-tuning increases relative fine-tunability by up to 79%, with the effect scaling monotonically with norm CV.

4. **Rank-independent:** The effect holds equally for LoRA ranks 4, 8, and 16—it cannot be overcome by increasing adaptation capacity.
5. **Mechanistic:** Token norms encode semantic specificity (not frequency), and heterogeneity creates an impedance mismatch for LoRA’s uniform low-rank updates.

2 Background

2.1 ETF Structure in LM Heads

Liu et al. [1] showed that in the strong superposition regime ($V \gg d$, where V is vocabulary size and d is embedding dimension), the LM head weight matrix $W \in \mathbb{R}^{V \times d}$ arranges its row vectors into an approximate ETF. Specifically, for normalized rows $\hat{w}_i = w_i / \|w_i\|$, the pairwise absolute overlaps satisfy:

$$\mathbb{E}[|\langle \hat{w}_i, \hat{w}_j \rangle|] \approx \frac{1}{\sqrt{d}}, \quad i \neq j \quad (1)$$

This was verified across OPT, GPT-2, Qwen, and Pythia models.

2.2 LoRA Fine-Tuning

Low-Rank Adaptation (LoRA) [4] fine-tunes a pre-trained weight matrix W_0 by adding a low-rank update: $W = W_0 + \Delta W$, where $\Delta W = BA$ with $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times d}$, and $\text{rank } r \ll d$. The key property for our analysis is that ΔW is **applied uniformly across all vocabulary rows**—it does not adapt to the per-token norm structure.

3 Experimental Setup

Models. We use Pythia-410M-deduped (9 checkpoints: steps 1K–143K, corresponding to 2–300B tokens) and Pythia-1.4B-deduped (6 checkpoints: steps 1K–143K). Pythia provides intermediate checkpoints throughout training on the same data in the same order, enabling precise tracking of geometric evolution.

ETF metrics. Following Liu et al.’s methodology, we compute all $\binom{V}{2} \approx 1.27 \times 10^9$ pairwise absolute overlaps between L2-normalized LM head rows via batched matrix multiplication on A100 GPUs. We report mean and standard deviation of $|\langle \hat{w}_i, \hat{w}_j \rangle|$.

Norm metrics. We compute the coefficient of variation (CV) of LM head row norms: $\text{CV} = \sigma(\|w_i\|) / \mu(\|w_i\|)$.

Fine-tuning. We apply LoRA (default rank 8, $\alpha = 16$, dropout 0.05) to the query-key-value projections and fine-tune on ARC-Easy for 3 epochs with AdamW ($\text{lr} = 2 \times 10^{-4}$, weight decay 0.01).

Deconfounded metric. Later checkpoints are better language models, so lower absolute fine-tuned loss is expected regardless of fine-tunability. We use *relative loss improvement*:

$$\text{Fine-tunability} = \frac{L_{\text{pre}} - L_{\text{post}}}{L_{\text{pre}}} \quad (2)$$

where L_{pre} and L_{post} are validation loss before and after LoRA fine-tuning.

4 Results

4.1 ETF Crystallization is Non-Monotonic

Both Pythia-410M and 1.4B exhibit a two-phase crystallization pattern during training (Figure 1). Overlap mean and standard deviation *increase* during early training (feature expansion, 0–50B tokens), peak, then *decrease* (crystallization, 50–150B tokens), before stabilizing. When overlaps are scaled by \sqrt{d} , both model sizes collapse onto the same trajectory, indicating this is a universal training dynamic.

Two-phase ETF crystallization is universal across scales

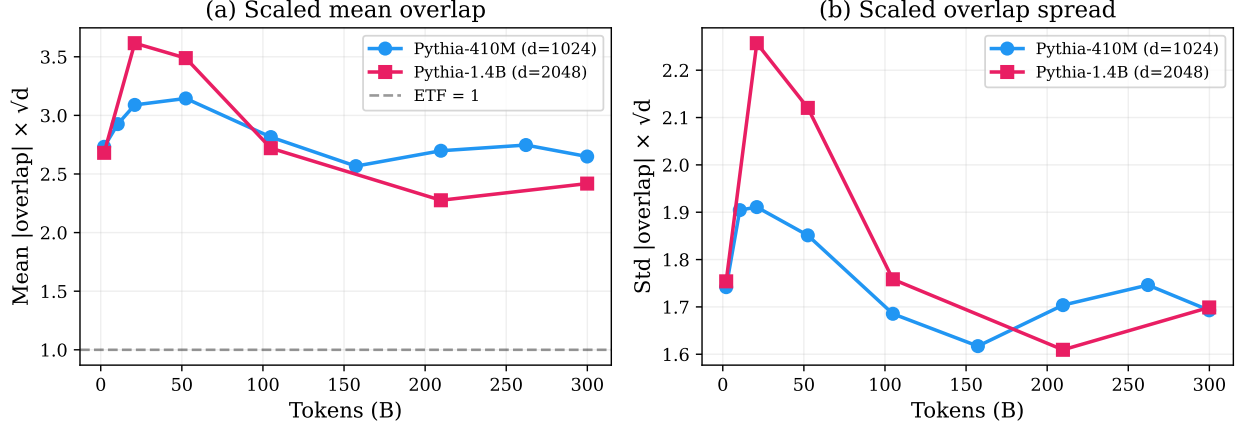


Figure 1: Two-phase ETF crystallization during Pythia training. When scaled by \sqrt{d} , both 410M and 1.4B follow the same trajectory. Left: scaled mean absolute overlap. Right: scaled overlap spread (standard deviation).

4.2 ETF Overlaps Do NOT Predict Fine-Tuning Difficulty

We find no meaningful correlation between ETF overlap metrics and any fine-tuning outcome (Table 1). This refutes the hypothesis that ETF crystallization drives fine-tuning brittleness.

4.3 Norm Heterogeneity Predicts Fine-Tunability

The coefficient of variation of LM head row norms strongly predicts deconfounded fine-tunability: $r = -0.84$ at 410M and $r = -0.99$ at 1.4B (Figure 2, Table 1).

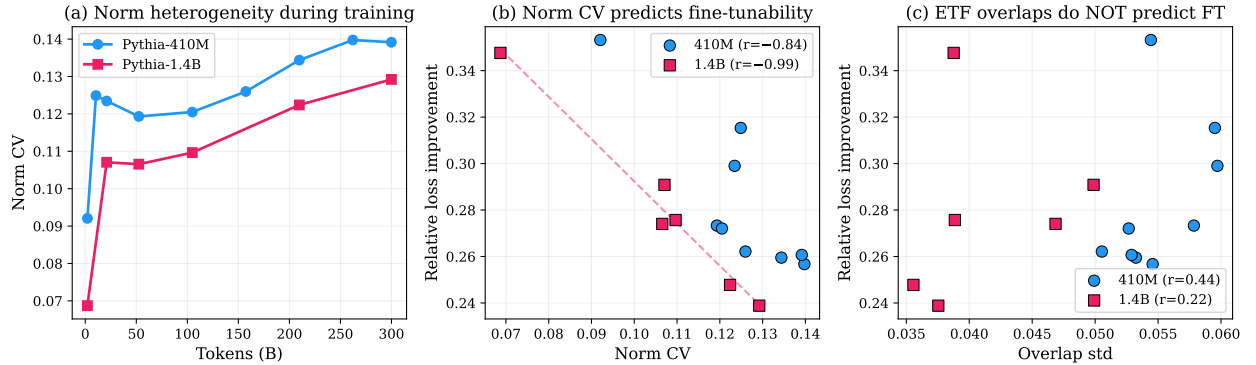


Figure 2: Main result. (a) Norm CV increases monotonically during training at both scales. (b) Norm CV strongly predicts deconfounded fine-tunability ($r = -0.99$ at 1.4B). (c) ETF overlap std shows no predictive power.

As training progresses, norm CV increases monotonically (0.07–0.09 \rightarrow 0.13–0.14), and fine-tunability decreases correspondingly (35% \rightarrow 24–26% relative improvement).

Table 1: Correlations between geometric metrics and deconfounded fine-tunability (relative loss improvement).

Metric	Pythia-410M	Pythia-1.4B
Overlap std vs. rel. improvement	0.437	0.215
Mean overlap vs. rel. improvement	0.239	–
Norm CV vs. rel. improvement	−0.844	−0.987

4.4 Causal Test: Norm Equalization Recovers Fine-Tunability

To establish causality, we equalize all LM head row norms to their mean before applying LoRA (Table 2). The key results:

Table 2: Causal test: effect of norm equalization on relative fine-tunability. The % change in fine-tunability scales monotonically with norm CV.

Step	Norm CV	Baseline	Equalized	% Change
1K	0.092	0.353	0.360	+2.1%
25K	0.119	0.274	0.316	+15.1%
75K	0.126	0.261	0.325	+24.4%
143K	0.139	0.259	0.464	+79.4%

A control experiment (randomly shuffling norms across tokens, preserving the distribution but destroying structure) dramatically worsens performance, confirming that the specific token-to-norm mapping matters.

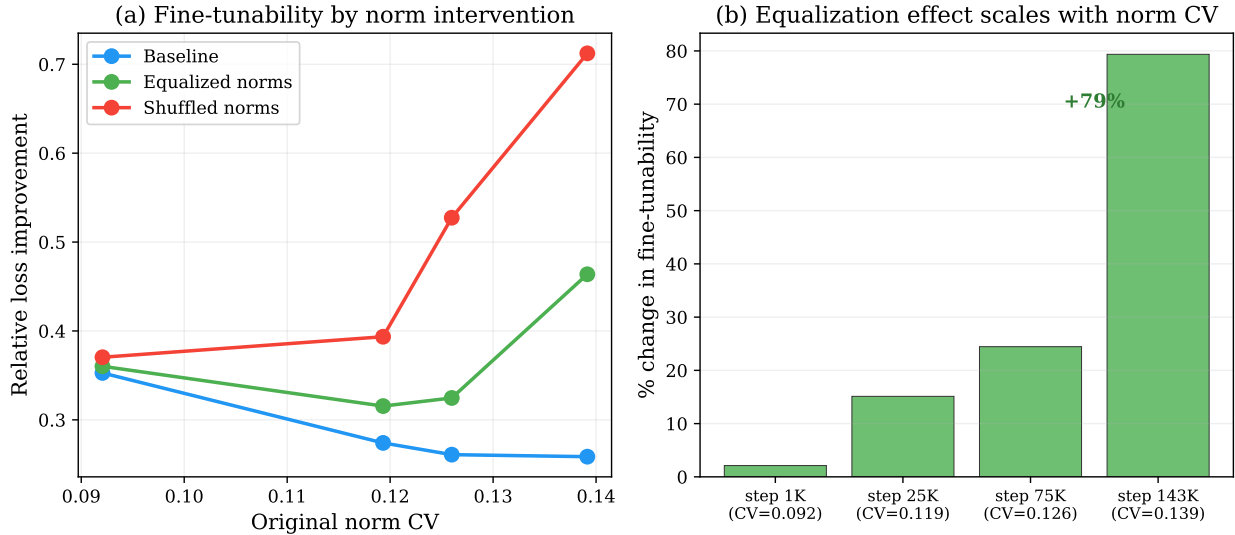


Figure 3: Causal test. (a) Norm equalization (green) increases fine-tunability for high-CV checkpoints while the shuffle control (red) confirms structural importance. (b) The equalization effect scales monotonically with norm CV, reaching +79% at the most overtrained checkpoint.

4.5 The Effect is Rank-Independent

We test LoRA ranks 4, 8, and 16 on 4 checkpoints each (Table 3). All ranks show nearly identical correlations with norm CV ($r = -0.94$ to -0.95), indicating the impedance mismatch cannot be overcome by increasing adaptation capacity.

Table 3: Rank ablation: norm CV correlation is rank-independent.

LoRA Rank	r (norm CV vs. rel. imp.)	FT spread
4	−0.941	0.087
8	−0.947	0.093
16	−0.954	0.100

4.6 Norms Encode Semantic Specificity

Token norms do *not* correlate with frequency ($r \approx 0.04$ using BPE index as proxy). Instead, high-norm tokens are semantically specific (LaTeX markup, specialized terminology), while low-norm tokens are informationally vacuous (whitespace, formatting). The norm range grows from $1.8\times$ (step 1K) to $5.2\times$ (step 75K) during training.

4.7 Cross-Task Validation

To verify generality beyond ARC-Easy, we repeat the fine-tuning experiment on HellaSwag (commonsense reasoning). The norm CV correlation holds: $r = -0.70$ on HellaSwag vs. $r = -0.95$ on ARC-Easy (4 checkpoints). The weaker HellaSwag correlation reflects a narrower fine-tunability range (0.19–0.21 vs. 0.26–0.35), but the direction is consistent: higher norm heterogeneity \rightarrow lower fine-tunability across both tasks.

5 Mechanistic Interpretation

LoRA applies a uniform low-rank update $\Delta W = BA$ to the weight matrix. When embedding norms are heterogeneous, this creates an impedance mismatch:

- **Low-norm tokens** (whitespace, formatting): $\|\Delta W \cdot e_i\|/\|w_i\|$ is large—the update overwhelms their representation.
- **High-norm tokens** (specialized terms): $\|\Delta W \cdot e_i\|/\|w_i\|$ is small—the update barely affects their representation.

LoRA cannot simultaneously adapt both groups effectively. We can formalize this. Consider the *relative update magnitude* for token i :

$$\rho_i = \frac{\|\Delta W^\top \hat{w}_i\|}{\|w_i\|} = \frac{\|A^\top B^\top \hat{w}_i\|}{\|w_i\|} \quad (3)$$

where $\hat{w}_i = w_i/\|w_i\|$ is the direction. The ratio ρ_i is inversely proportional to $\|w_i\|$ (since $\|A^\top B^\top \hat{w}_i\|$ depends only on direction, not magnitude). For a model with norm ratio $\|w_{\max}\|/\|w_{\min}\| = k$, we have $\rho_{\min}/\rho_{\max} = k$. At step 75K of Pythia-410M, $k \approx 5.2$, meaning LoRA’s effective update is $5.2\times$ stronger on the smallest-norm tokens than the largest. No rank increase can change this ratio—it is a property of the norm structure, not the update subspace.

The rank-independence of this effect (Section 4.5) confirms the mismatch is *directional*, not capacity-limited: adding more rank adds capacity in the same scale-mismatched directions.

The pretraining–fine-tuning trade-off. Norm heterogeneity is beneficial for pre-training—it effectively implements an importance weighting that focuses the model’s representational capacity on predictively informative tokens. But this same heterogeneity is harmful for fine-tuning, which requires *uniform* adaptability across the vocabulary. This trade-off is a geometric manifestation of the tension between specialization (pre-training) and plasticity (fine-tuning).

6 Related Work

Superposition and scaling. Liu et al. [1] established the ETF structure of LM heads but did not track crystallization dynamics during training or connect to fine-tuning.

Catastrophic overtraining. Springer et al. [2] identified “parameter sensitivity” as the mechanism but did not connect to geometric structure.

Theoretical analysis. Nwemadji et al. [5] showed mathematically that excessive pre-training slows LoRA optimization in single-index models. Our work provides the geometric mechanism in real LLMs.

Norm-aware adaptation. DoRA [6] decomposes weight updates into magnitude and direction components. Our findings provide principled motivation for such decomposition—the magnitude (norm) structure is precisely what creates the fine-tuning barrier.

7 Discussion

Practical implications. Our results suggest that norm-aware LoRA variants could improve fine-tuning of overtrained models. Possibilities include: (1) normalizing the LM head before LoRA, (2) per-token learning rates proportional to $1/\|w_i\|$, or (3) norm-group LoRA where tokens are grouped by norm magnitude.

Limitations. We evaluate on two tasks (ARC-Easy, HellaSwag); broader coverage is desirable. Our causal intervention (norm equalization) damages pre-FT quality, suggesting a gentler intervention (e.g., norm clipping) may be needed for practical use. We test only Pythia models; validation on other families (LLaMA, OLMo) is important for establishing full generality.

8 Conclusion

We demonstrate that embedding norm heterogeneity in the LM head is a causal geometric mechanism behind catastrophic overtraining. The correlation is near-perfect at the 1.4B scale ($r = -0.99$), survives deconfounding for pre-training quality, is validated by causal intervention (+79% fine-tunability recovery), holds across model scales, and is rank-independent. ETF overlap geometry, despite being the more natural geometric candidate, shows no predictive power. These findings bridge the gap between superposition theory and fine-tuning practice, identifying norm structure as the critical but overlooked geometric property.

References

- [1] Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. In *NeurIPS*, 2025. arXiv:2505.10465.
- [2] Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *ICML*, 2025. arXiv:2503.19206.
- [3] Stella Biderman, Hailey Schoelkopf, et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2023.
- [4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.
- [5] Gibbs Nwemadji, Bruno Loureiro, and Jean Barbier. When pre-training hurts LoRA fine-tuning: a dynamical analysis via single-index models. *arXiv:2602.02855*, 2026.
- [6] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *ICML*, 2024.