Tracing the Representation Geometry of Language Models from Pretraining to Post-training

A spectral view of learning phases and post-training trade-offs

Paper by M. Z. Li, K. K. Agrawal, A. Ghosh, K. K. Teru, A. Santoro, G. Lajoie, and B. A. Richards

Research group reading 25 October 2026

arXiv:2509.23024v1

Falling loss conceals changes in learned features

Research question

How do internal feature directions change during pretraining and post-training, and when do those changes align with model behavior?

Why spectra?

Loss averages prediction error. A covariance spectrum records how hidden-state variance is distributed across feature directions.

Checkpoint sequence $\longrightarrow$ feature geometry $\longrightarrow$ behavioral comparison

Paper, Sec. 1, PDF pp. 1–3; Fig. 1, PDF p. 2.

A checkpoint becomes a covariance spectrum

1. Feed the same evaluation sequences through each checkpoint.
2. Take the last-layer, last-token vector from each sequence.
3. Center the $M \times d$ feature matrix F; diagonalize its covariance.

$$F = \begin{bmatrix} f_\theta(x_1)^\top \\ \vdots \\ f_\theta(x_M)^\top \end{bmatrix}, \qquad \widehat{\Sigma} = \frac{1}{M} F^\top F, \qquad \lambda_1 \geq \cdots \geq \lambda_d \geq 0.$$

What the spectrum records

The relative variance along principal feature directions, rather than one aggregate prediction score.

Paper, Sec. 2.1, PDF p. 3; Fig. 1A–B, PDF p. 2.

RankMe and α_{ReQ} describe complementary geometry

$$p_i = \frac{\lambda_i}{\sum_j \lambda_j}, \qquad \mathrm{RankMe} = \exp\left(-\sum_i p_i \log p_i\right), \qquad \lambda_i \propto i^{-\alpha_{\mathrm{ReQ}}}.$$

Measure	Larger value means	Smaller value means
RankMe	variance spread over more axes	fewer effective axes
α_{ReQ}	steeper spectrum, more concentration	flatter spectrum

Read the pair together

Expansion: RankMe $\uparrow$, $\alpha_{\mathrm{ReQ}} \downarrow$. Anisotropic compression: RankMe $\downarrow$, $\alpha_{\mathrm{ReQ}} \uparrow$.

Paper, Sec. 2.1 and Eq. (1), PDF pp. 3–4.

The experiments span pretraining and adaptation

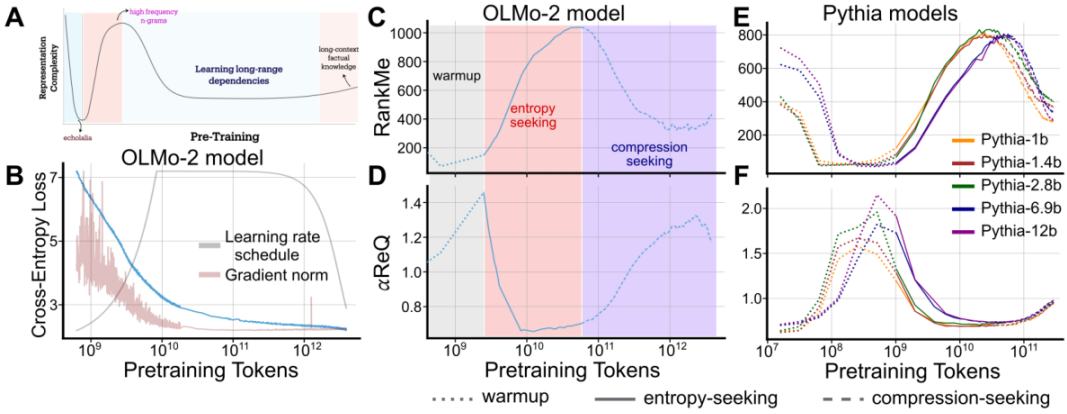
- **OLMo-2, 1B and 7B:** intermediate pretraining checkpoints; separate 1B SFT runs compare Anthropic-HH and AlpacaFarm.
- **Pythia, 70M–12B:** common Pile training recipe across scales; probe ∞-gram alignment, SciQ, and TriviaQA.
- **Tülu-3.1, 8B:** Llama-3.1 base $\rightarrow$ SFT $\rightarrow$ DPO $\rightarrow$ RLVR; probe geometry and AMC-23 pass@k.

Common geometry probe

FineWeb sequences; main spectral setup uses 15,000 sequences of at most 512 tokens.

Paper, Sec. 3, PDF pp. 5–6; Appendix A, PDF pp. 19–21. FineWeb is a geometry probe, not a shared pretraining corpus.

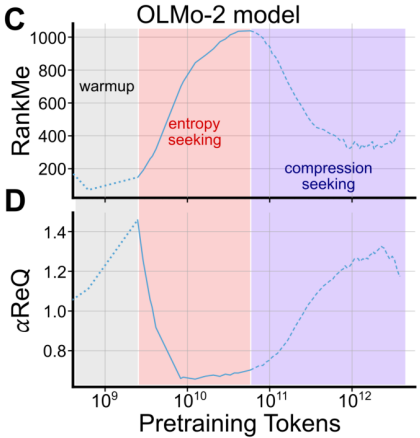
Pretraining passes through three geometric phases



Warmup collapse → entropy-seeking expansion → anisotropic compression

Paper, Fig. 2, PDF p. 5. Horizontal axes show pretraining tokens on log scales.

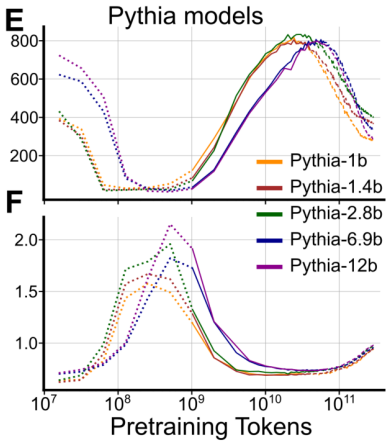
OLMo-2 geometry is non-monotonic



- Initial collapse: RankMe falls, α_{ReQ} rises.
- Middle phase: RankMe rises sharply as the spectrum flattens.
- Later phase: the spectrum becomes more concentrated.
- Very late OLMo checkpoints hint at another expansion.

Paper, Fig. 2C–D, PDF p. 5; Sec. 3.1, PDF pp. 6–7.

Pythia reproduces the pattern across scales



Paper, Fig. 2E–F, PDF p. 5; Appendix Fig. 11, PDF p. 30.

- The 1B–12B curves share collapse, expansion, and compression.
- Their phase timing shifts with model scale.
- Appendix results also show the pattern below 1B parameters.

The phase shape survives several measurement choices

Changed quantity	Reported outcome
Layer	Intermediate layers show related phase trajectories.
Sequence count and length	OLMo-2 phase shape persists under tested choices.
Probe dataset	Pythia and OLMo retain broadly similar phase shapes; absolute RankMe can vary.
Model size	Pythia models below 1B also show the sequence.

Practical use

The last layer is a compact tracking signal for the training trajectory in these model families.

Paper, Fig. 3, PDF p. 6; Appendix Figs. 8–11, PDF pp. 29–30.

∞-gram alignment probes corpus-local statistics

- An ∞-gram model backs off from the longest corpus-matched prefix.
- Mem_∞ is Spearman rank correlation between its answer likelihoods and the LLM's.
- On TriviaQA, 88.73% of matched suffixes have length ≤ 5.

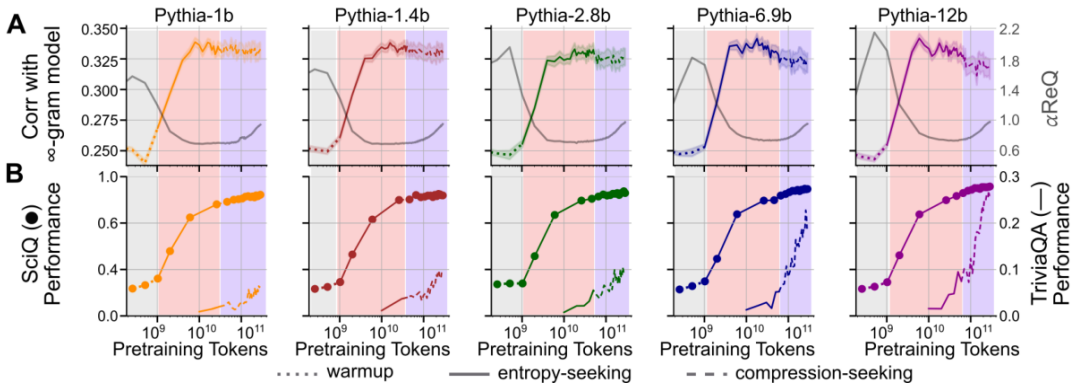
$$\mathrm{Mem}_\infty = \rho_s\big(\bar{P}_{\infty,\mathcal{D}}(y \mid u \oplus x), \bar{P}_{\mathrm{LLM}}(y \mid u \oplus x)\big)\,.$$

Interpretation

Higher correlation means closer alignment with corpus n-gram statistics, not measured verbatim copying.

Paper, Sec. 2.2, Eq. (2), PDF p. 4; Table 1, PDF p. 7.

Local-statistic alignment peaks before TriviaQA rises



∞-gram alignment rises in expansion; TriviaQA accuracy grows mostly during compression.

Paper, Fig. 4A–B, PDF p. 8. Each column is one Pythia scale; dots are SciQ, thin lines TriviaQA.

SciQ and spectrum co-vary across checkpoints

Model	corr(SciQ, α_{ReQ})	corr(SciQ, RankMe)
Pythia-1B	0.810	−0.759
Pythia-6.9B	0.837	−0.885
Pythia-12B	0.836	−0.839
OLMo-2-1B	0.540	−0.616

Reading the result

Better SciQ checkpoints tend to have steeper spectra and lower effective rank in this analysis. The comparison follows training trajectories, so it does not isolate a causal role for the spectrum.

Paper, Appendix Table 9, PDF p. 30. Reported coefficients; p-values are in the paper.

A steep spectrum does not make its tail dispensable

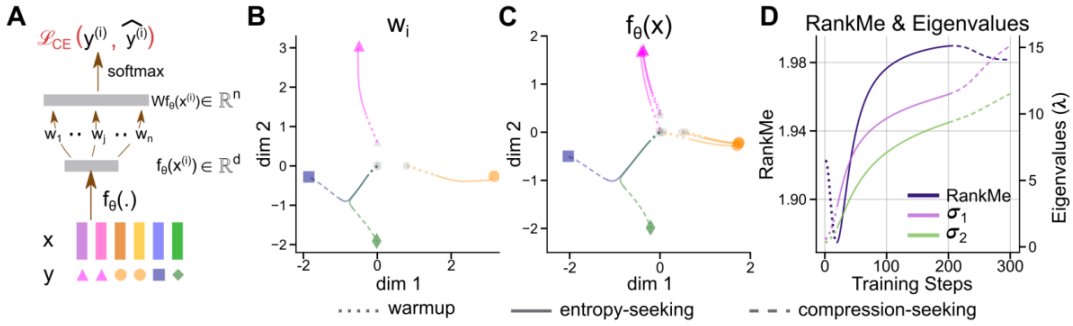
SciQ accuracy	Original	Keep top 10	Remove top 10
Pythia-1B	0.838	0.225	0.849
Pythia-6.9B	0.896	0.202	0.893
OLMo-2-7B	0.970	0.155	0.966

Implication

Compression changes the distribution of variance; useful information remains spread through the full spectrum.

Paper, Table 2, PDF p. 7; discussion, PDF p. 13.

A toy cross-entropy model reproduces the phase sequence



Skewed class frequencies + feature bottleneck $d < |\mathcal{V}|$ produce expansion followed by concentration.

Paper, Sec. 3.3 and Fig. 5, PDF pp. 9–10. The figure is a simplified linear model.

Gradient dynamics favor existing directions under assumptions

$$Z = FW, \quad A = \mathrm{softmax}(Z) - Y, \qquad \dot{F} = -\eta A W^\top, \quad \dot{W} = -\eta F^\top A.$$

1. These updates preserve $F^\top F - W W^\top$.
2. If that difference is zero at initialization, feature and classifier singular values stay paired.
3. For a paired singular value s_i: $\dot{s}_i = -\eta g_i s_i$; g_i depends on residuals and current vectors.

Scope

The derivation assumes linear features and orthonormal input rows. The sign of g_i changes with training; the equation alone does not force compression.

Paper, Appendix B.1–B.4, PDF pp. 23–27; Eqs. (12), (15)–(20).

Toy controls locate the ingredients for compression

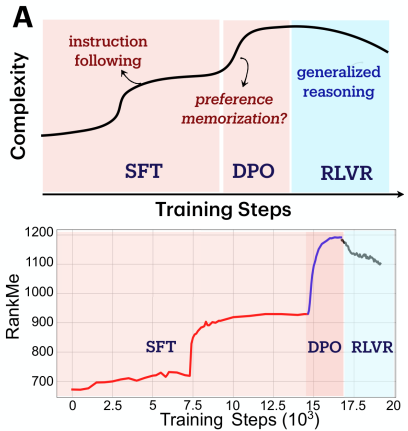
Toy-model intervention	Observed rank trajectory
Uniform class frequencies	Expansion continues; no analogous compression.
Remove feature bottleneck	Expansion continues; no analogous compression.
Replace cross-entropy with MSE	Rank expands and saturates.

Mechanistic reading

Frequent classes separate first; later classes reuse limited feature directions, causing anisotropic growth.

Paper, Sec. 3.3, PDF pp. 9–10; Appendix Figs. 12–13, PDF pp. 31–32.

Post-training objectives move geometry in opposite directions



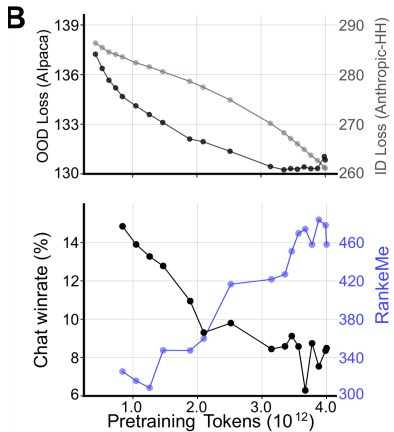
- SFT: RankMe rises.
- DPO: RankMe rises sharply; α_{ReQ} falls.
- RLVR: RankMe falls.

Setup

Sequential Tülu-3.1 8B checkpoints derived from Llama-3.1-8B.

Paper, Sec. 3.4 and Fig. 6A, PDF pp. 10–11.

SFT becomes more sensitive to its instruction data



- OLMo-2-1B is fine-tuned on Anthropic-HH; AlpacaFarm tests distribution shift.
- ID loss keeps improving; OOD loss falls, then levels off or slightly rebounds.
- Anthropic-HH-SFT vs AlpacaFarm-SFT chat win rate drops from about 14% to 9%.

Paper, Fig. 6B, PDF p. 11; Appendix Fig. 14, PDF p. 33. The body text says OOD loss "increases," but the plots/captions support saturation after an overall fall.

DPO's pairwise objective accompanies rank expansion

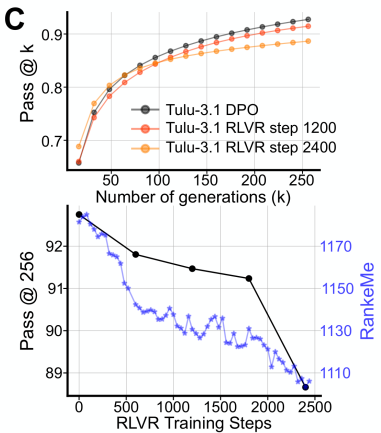
$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E} \log \sigma(\widehat{r}_\theta(x, y_w) - \widehat{r}_\theta(x, y_l)),$$

$$\widehat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}.$$

- y_w and y_l are preferred and dispreferred responses.
- The loss can be written as a two-way contrastive softmax.
- In Tülu checkpoints, RankMe rises and α_{ReQ} falls during DPO.

Paper, Eq. (3), PDF p. 4; Sec. 3.4 and Eq. (5), PDF p. 11. The contrastive analogy is an interpretation.

RLVR gains at small k but loses at large k



- At AMC-23 pass@16, the late RLVR checkpoint beats post-DPO.
- At pass@256, post-DPO and an intermediate checkpoint beat late RLVR.
- RankMe falls during the same RLVR stage.

Reading

Reward-aligned solutions become easier to sample, while high-k coverage narrows.

Paper, Fig. 6C, PDF p. 11; pass@k estimator with $N = 512$, Eq. (4), PDF p. 5.

Geometry is a useful trajectory marker, with clear limits

1. RankMe and α_{ReQ} reveal changes hidden by average loss.
2. Expansion aligns with ∞-gram statistics; compression aligns with later QA gains and RLVR behavior.
3. The main behavior links are correlational; the mechanistic derivation is for a linear toy model.

Empirical scope

English-language model families up to 12B parameters; spectral estimates need roughly 10^4 probe sequences.

Paper, Secs. 3–6, PDF pp. 5–14.

A useful next test: intervene on geometry

Question for discussion

Does the compression phase help generalization, or does it merely accompany another training change?

1. Match checkpoints on loss and compute budget.
2. Perturb head and tail eigendirections separately.
3. Vary the required context length and measure task performance.

This is a proposed experiment, not a result reported in the paper.

Motivated by Paper, Fig. 4 and Table 2, PDF pp. 7–8; limitations, PDF p. 13.

Backup: key measurement and SFT settings

Setting	Paper value
Main spectral probe	FineWeb sample-10BT; 15,000 sequences; max 512 tokens; batch 16
OLMo-2-1B SFT	Anthropic-HH or AlpacaFarm; 2 epochs; learning rate 10^{-5}; 5 seeds
ID/OOD loss evaluation	10,000 sequences; max 1,024 tokens
RLVR pass@k	AMC-23; $N = 512$ samples; $k \leq 256$

Paper, Appendix A, Tables 5–8, PDF pp. 21–22; Eq. (4), PDF p. 5.

Backup: full top-direction ablation

Model	Original	Top 10 out	Top 10 only	Top 50 out	Top 50 only
Pythia-1B	.838	.849	.225	.835	.318
Pythia-1.4B	.866	.855	.232	.859	.324
Pythia-2.8B	.884	.880	.219	.873	.317
Pythia-6.9B	.896	.893	.202	.906	.327
OLMo-2-1B	.953	.943	.199	.954	.326
OLMo-2-7B	.970	.966	.155	.970	.308

Paper, Table 2, PDF p. 7. SciQ accuracy after retaining or removing leading eigendirections.