

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

Melody Zixuan Li^{1,2,*}, Kumar Krishna Agrawal^{3,*}, Arna Ghosh^{1,2,9,*}, Komal Kumar Teru⁴, Adam Santoro^{5,†},
Guillaume Lajoie^{2,6,9} and Blake A. Richards^{1,2,7,8,9}

¹Computer Science, McGill University, ²Mila - Quebec AI Institute, ³UC Berkeley, ⁴Cohere, ⁵Google Deepmind, ⁶Mathematics and Statistics, Université de Montréal, ⁷Neurology & Neurosurgery and Montreal Neurological Institute, McGill University, ⁸CIFAR Learning in Machines & Brains Program, ⁹Google, Paradigms of Intelligence Team, *Equal contribution, †Advisory capacity only

Abstract: Standard training metrics like loss fail to explain the emergence of complex capabilities in large language models. We take a spectral approach to investigate the geometry of learned representations across pretraining and post-training, measuring effective rank (RankMe) and eigenspectrum decay (α_{ReQ}). With OLMo (1B-7B) and Pythia (160M-12B) models, we uncover a consistent non-monotonic sequence of three geometric phases during autoregressive pretraining. The initial “warmup” phase exhibits rapid representational collapse. This is followed by an “entropy-seeking” phase, where the manifold’s dimensionality expands substantially, coinciding with peak n-gram memorization. Subsequently, a “compression-seeking” phase imposes anisotropic consolidation, selectively preserving variance along dominant eigendirections while contracting others, a transition marked with significant improvement in downstream task performance. We show these phases can emerge from a fundamental interplay of cross-entropy optimization under skewed token frequencies and representational bottlenecks ($d \ll |\mathcal{V}|$). Post-training further transforms geometry: SFT and DPO drive “entropy-seeking” dynamics to integrate specific instructional or preferential data, improving in-distribution performance while degrading out-of-distribution robustness. Conversely, RLVR induces “compression-seeking”, enhancing reward alignment but reducing generation diversity.

1. Introduction

Loss curves during training offer an incomplete account of how large language models (LLMs) learn specific behaviors (Wei et al., 2022; Ganguli et al., 2022). While training loss decreases monotonically (Kaplan et al., 2020; Hoffmann et al., 2022), model capabilities and internal representational structures exhibit significant qualitative shifts (Singh et al., 2023; Brown et al., 2023; Singh et al., 2024). This disconnect highlights a fundamental challenge: How do high-dimensional distributed representations within LLMs evolve during training, and how do these representational transformations give rise to emergent capabilities?

We answer this question by using spectral analysis to quantify the geometric evolution of LLM representations. We discover that this evolution is not a smooth progression but a consistent, three-phase dynamic. Our method centers on the spectral properties of the covariance matrix of last-token representations, which capture rich information about the model’s internal representations, especially when using causal attention. To measure this geometric structure, we compute two metrics from the eigenspectrum of these matrices: the effective rank (RankMe), derived from the Von Neumann entropy, and the power-law decay rate (α_{ReQ}) of the eigenvalues (Garrido et al., 2023; Agrawal et al., 2022). These spectral measures of representation geometry have been linked theoretically and experimentally to generalization in down-

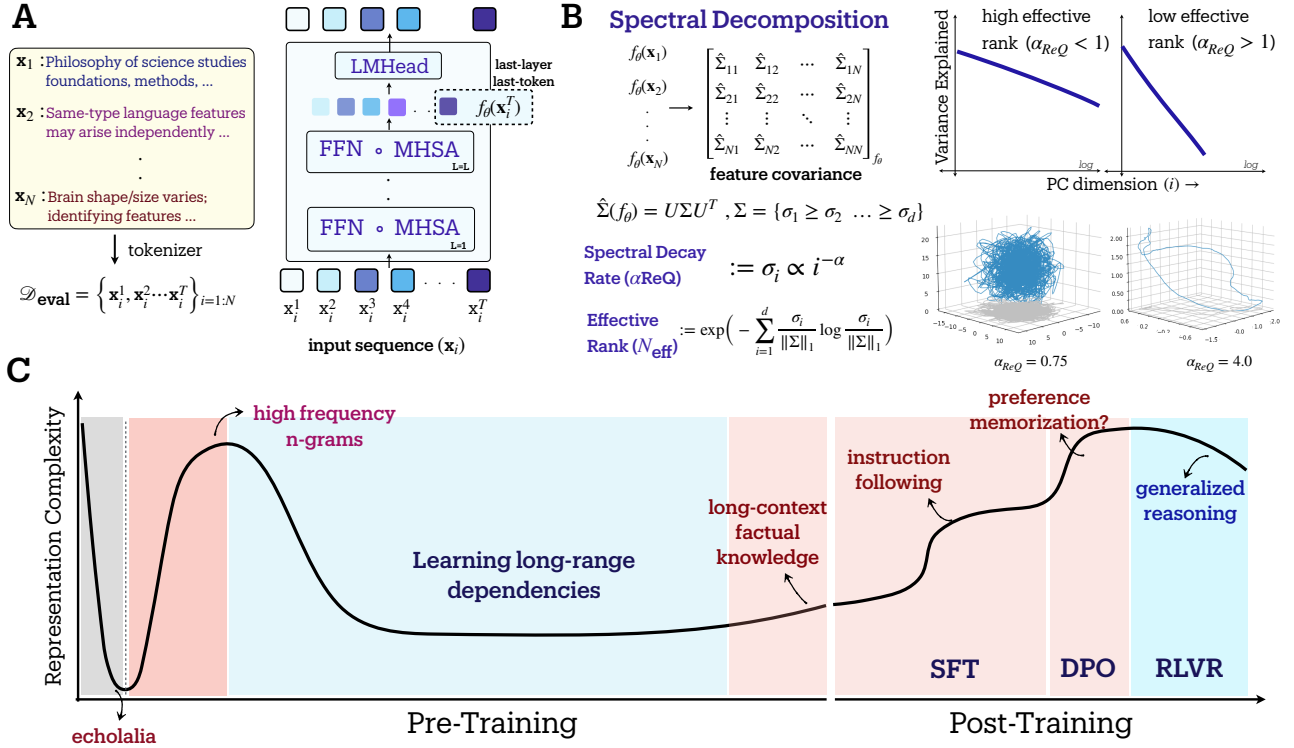


Figure 1: Spectral framework reveals three universal phases in LLM training. (A) LLM representations analyzed via empirical feature covariance $\hat{\Sigma}(f_\theta)$ of last-token hidden states $f_\theta(x_i)$. (B) Two complementary spectral metrics: α_{ReQ} measures eigenspectrum decay rate (variance concentration), while RankMe quantifies effective rank (utilized dimensionality). (C) Pretraining exhibits three phases: “warmup” (rapid collapse), “entropy-seeking” ($2\text{-}3\times$ expansion coinciding with n-gram memorization), and “compression-seeking” (anisotropic consolidation enabling long-context understanding). Post-training continues these dynamics: SFT/DPO induce “entropy-seeking” while RLVR induces “compression-seeking”.

stream tasks (Bartlett et al., 2020; Thilak et al., 2023). Intuitively, representation geometry tells us about the model’s expressive capacity, utilization, and amount of data compression.

Our analysis shows that LLM pretraining unfolds through a consistent sequence of distinct geometric phases marked by non-monotonic evolution of spectral properties. These phases correlate with significant shifts in the model’s expressive power and information compression ability (c.f. Figure 1):

- An initial “warmup” phase, coinciding with learning rate ramp-up, where there is a rapid collapse of representations onto dominant data manifold directions.
- An “entropy-seeking” phase marked by manifold expansion in many directions, which correlates with an increase in n-gram distributional memorization.
- A “compression-seeking” phase with anisotropic consolidation along principal feature eigenvectors shows enhanced learning of long-range dependencies and robust generalization.

We further develop mechanistic insights from analytically tractable toy models, demonstrating that these geometric phase transitions are influenced by the interplay of cross-entropy optimization, information bottlenecks, and skewed data distributions.

Our investigation of post-training stages reveals analogous geometric shifts: Supervised fine-tuning (SFT) continues an “**entropy-seeking**”-like manifold expansion with concomitant assimilation of specific instructions. Reinforcement Learning from Verifiable Rewards (RLVR) produces a “**compression-seeking**”-like contraction, which can consolidate reward-aligned behaviors yet curtail generative novelty and exploration. These findings offer a more granular view of LLM training, and offer some practical implications for optimizing LLM training and adaptation pipelines based on desired downstream outcomes.

2. Methods

2.1. Spectral Analysis, Matrix Entropy, and Effective Rank

Last token representations in autoregressive language models: A rigorous understanding of LLM capabilities necessitates a precise characterization of the *geometry of their learned representations*. An autoregressive language model processes an input sequence of discrete tokens $\mathbf{s} = (t_1, t_2, \dots, t_N)$, transforming each token t_k through its l layers (conditioned on preceding tokens $t_{<k}$) into a sequence of high-dimensional continuous vectors $\mathbf{f}_\theta^{(l)}(t_k|t_{<k})$. For autoregressive models, the representation of the final token (t_N) at the last layer, $\mathbf{y}_N := \mathbf{f}_\theta^{(L)}(t_N|t_{<N})$, is particularly pivotal. Its significance stems from different factors: (i) it directly parameterizes the predictive distribution for the subsequent tokens $P(t_{N+1}|t_1, \dots, t_N)$; (ii) it synthesizes information from the entire context $t_{\leq N}$ to inform this prediction, meaning it inherently reflects the model’s capacity for contextual understanding; and (iii) is often used as input to task-specific layers in downstream applications.

High-dimensional representation complexity metrics: To quantitatively measure representation geometry, we perform spectral analysis of the feature covariance matrix. Given a set of M input sequences, we form a feature matrix $\mathbf{F} \in \mathbb{R}^{M \times d}$, each row is a feature vector of the last token \mathbf{y}_N for each input. Assuming the features are centered, the empirical covariance matrix is $\hat{\Sigma} := \frac{1}{M} \mathbf{F}^T \mathbf{F}$. The eigenspectrum of $\hat{\Sigma}$, denoted by eigenvalues $\{\sigma_i(\hat{\Sigma})\}_{i=1}^d$, measures the concentration of information along the principal axes of variation. The distribution of $\{\sigma_i\}_{i=1}^d$ provides a quantitative description of feature geometry: a sharp decay indicates information compressed in a lower-dimensional subspace (anisotropic geometry), while a slow decay indicates a high-dimensional subspace is utilized.

This spectral perspective motivates using *matrix entropy* to measure the uniformity of the eigenvalue distribution. If $p_i = \sigma_i / (\sum_j \sigma_j)$ is the proportion of variance along the i -th principal axis, the Von Neumann *entropy-based effective rank* (Roy and Vetterli, 2007; Garrido et al., 2023) is defined as:

$$\text{RankMe} := \exp \left(S(\hat{\Sigma}) \right) = \exp \left(- \sum_{i=1}^d p_i \ln p_i \right) \in (0, d]. \quad (1)$$

Low entropy indicates a skewed eigenvalue distribution, i.e. low-dimensional (anisotropic) representations, while high entropy implies a uniform spread, i.e. high-dimensional (isotropic) representations.

Our empirical studies also show that LLM activation matrices exhibit *heavy-tailed* eigenvalue spectra, i.e., a power law distribution where $\sigma_i \propto i^{-\alpha_{\text{ReQ}}}$, where $\alpha_{\text{ReQ}} > 0$ (Ghosh et al., 2022). Slower decay or smaller α_{ReQ} implies a more uniform spread of σ_i ’s (higher dimensional), and thus higher $S(\hat{\Sigma})$ and RankMe. Conversely, faster decay or larger α_{ReQ} implies representations are compactly packed along fewer principal directions (Stringer et al., 2019; Agrawal et al., 2022), yielding lower entropy and smaller