THE LOW-RANK CANONICAL GEOMETRY OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models predict the next token from vocabularies exceeding 10^5 tokens, suggesting an extremely high-dimensional output space. Yet it remains unclear whether their predictive distributions exploit this enormous space, or instead vary within a much smaller structure as context varies. We develop a spectral framework that characterizes this geometry directly in probability space. By separating the model's predictive marginal from context-dependent probability variation, the framework decomposes LLM next-token prediction into orthogonal token-side canonical modes, revealing both the dimensionality and orientation of the predictive geometry. Surprisingly, we find that LLM prediction is organized around a much smaller canonical geometry despite their vast vocabulary. Across model families and context distributions, nearly all context-dependent predictive variation is captured by a small set of dominant token modes: only 9.7%–24.0% of the token-space dimensions account for 99.9% of this variation. More surprisingly, this geometry remains largely intact under downstream adaptation. Fine-tuning changes the model's conditional predictions while largely preserving the concentration of the canonical spectrum and the dominant canonical token subspace, even across different downstream tasks. These observations motivate the *Canonical Subspace Hypothesis*: pretrained language models organize next-token prediction around a low-dimensional canonical token geometry that downstream adaptation largely reuses rather than reconstructs. More broadly, our framework provides a general tool for studying how the geometry of LLM prediction is formed, preserved, and transformed throughout training.

1 INTRODUCTION

The dimensional structure of a language model's predictions determines how its output distribution varies with context. Modern LLMs predict over vocabularies containing tens or hundreds of thousands of tokens (Grattafiori et al., 2024; Bai et al., 2023), yet it remains unclear whether their next-token distributions explore this enormous output space or vary within a much lower-dimensional structure. Prior work has studied the expressive capacity of next-token prediction, including architectural limitations imposed by the softmax output layer (Yang et al., 2017), the capacity of transformers to realize context-to-distribution mappings (Madden et al., 2025), and, more recently, which target next-token distributions can be elicited from pretrained language models (Wang et al., 2025). While these works characterize what distributions a model can represent or produce, they leave open the geometry of the distributions LLMs generate as context changes. Understanding this geometry may also shed light on why substantial behavioral adaptation can often be achieved through comparatively low-dimensional modifications (Aghajanyan et al., 2021; Hu et al., 2021).

Low-dimensional structure has been studied extensively in language models, but primarily in spaces other than the predictive distribution itself. Fine-tuning updates can occupy low-dimensional parameter subspaces (Aghajanyan et al., 2021; Ruppik et al., 2026); analyses of hidden states have identified structured and partially stable internal representations under adaptation (Jain et al., 2024; Confavreux et al., 2026); and the finite-dimensional hidden representation imposes rank constraints on pre-softmax logits (Yang et al., 2017; Godey & Artzi, 2026; Godey et al., 2024). More recently, low-rank structure has also been studied in logit representations spanning histories and future continuations (Golowich et al., 2026). These results characterize parameters, internal representations, or pre-softmax scores, but do not determine the geometry of the normalized probability distributions produced by the model.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

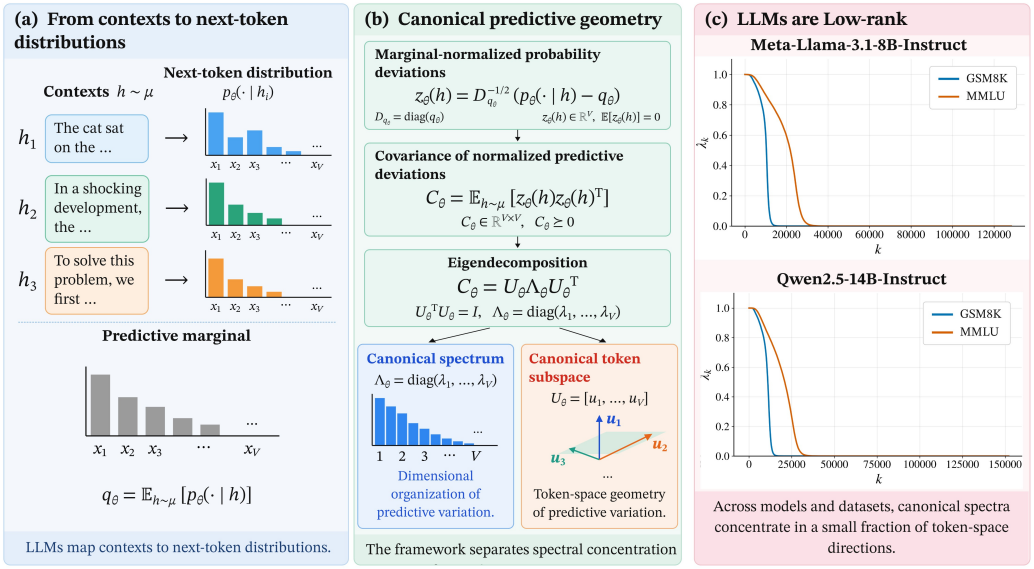


Figure 1: **Canonical geometry of LLM next-token prediction.** **(a)** An LLM maps each context h to a next-token distribution $p_\theta(\cdot \mid h)$, whose average over contexts is the predictive marginal q_θ. **(b)** Our framework represents context-dependent predictive variation by the marginal-normalized deviations $z_\theta(h) = D_{q_\theta}^{-1/2}(p_\theta(\cdot \mid h) - q_\theta)$ and eigendecomposes their covariance $C_\theta = U_\theta \Lambda_\theta U_\theta^\top$, yielding both a canonical spectrum Λ_θ and a canonical token subspace U_θ. **(c)** Applying this framework to LLMs reveals a strongly concentrated canonical spectrum: most context-dependent predictive variation lies in a small fraction of the available token-space directions.

An autoregressive language model maps each context h to a next-token distribution $p_\theta(\cdot \mid h) \in \Delta^{V-1}$. As the context changes, these distributions vary in a high-dimensional probability space. The central question we study is whether this variation spans a high-dimensional space or is concentrated along a much smaller set of recurring token-space directions. This question cannot be answered directly from low-rank structure in logits as in existing work Golowich et al. (2026): the softmax transformation is nonlinear, and variation in probability space is strongly affected by the highly non-uniform marginal frequencies of different tokens. A direct analysis therefore requires a representation defined at the level of the predictive distributions themselves.

To address this, we develop a spectral framework to characterize the geometry of LLM next-token prediction. As illustrated in Figure 1, we separate the model's context-independent predictive marginal from its context-dependent predictive variation, representing the resulting deviations in a marginal-normalized probability space. This construction yields a token-side canonical operator whose eigendecomposition identifies orthogonal directions of predictive variation. Its eigenvalues quantify the strength of variation along each direction across contexts, while its eigenvectors identify the corresponding token-space directions. The framework therefore characterizes two complementary properties of prediction: the *dimensional organization* of predictive variation and the *geometry* of the directions that carry it.

Our framework reveals a pronounced low-dimensional structure in LLMs' next-token distributions. **Most context-dependent variation in LLM next-token distributions is concentrated in a comparatively small set of dominant canonical token modes.** Figure 1 illustrates this behavior for Meta-Llama-3.1-8B-Instruct and Qwen2.5-14B-Instruct. Although their vocabularies contain 128,256 and 152,064 tokens, respectively, the canonical spectra concentrate nearly all predictive variation within a restricted portion of the available token space. The same spectral concentration is observed when the geometry is estimated independently from GSM8K and MMLU generations.

Beyond characterizing the dimensionality of next-token prediction, our framework allows us to study how its geometry changes under downstream adaptation. We surprisingly find that the low-dimensional spectral structure of $p_\theta(\cdot \mid h)$ is largely preserved during fine-tuning. Furthermore, the dominant token-side subspace itself remains strongly aligned with that of the foundation model,

including across different downstream tasks. These results suggest that fine-tuning largely reuses an existing predictive geometry rather than constructing a new one. By explicitly identifying and tracking the dominant predictive subspace of $p_\theta(\cdot \mid h)$, our framework provides a concrete geometric view of adaptation efficiency: substantial behavioral change can occur while reusing a compact predictive structure already established by the foundation model.

These observations motivate the **Canonical Subspace Hypothesis**: *language-model pretraining*[1] *organizes next-token prediction around a low-dimensional token geometry that downstream adaptation largely reuses rather than reconstructs*. This provides a predictive-space counterpart to recent evidence that fine-tuning can elicit, reorganize, or amplify structure already present in the base model (Jain et al., 2024; Yue et al., 2026; Confavreux et al., 2026).

Our contributions are threefold. First, we develop a spectral framework for characterizing LLM prediction directly in probability space, separating the strength of context-dependent predictive variation from the geometry of the token directions that carry it. Second, using this framework, we uncover a pronounced low-dimensional predictive geometry across model families and context distributions. Third, we study downstream fine-tuning with our framework and find that it largely preserves both the concentration of the canonical spectrum and the dominant canonical token subspace. Together, these results indicate that the ambient dimensionality of the vocabulary substantially overstates the dimensionality of the predictive variation used by an LLM.

2 Canonical Geometry of Next-Token Prediction

2.1 Canonical Decomposition and Predictive Dependence

An autoregressive language model assigns a next-token distribution $p_\theta(x \mid h)$ to each context $h = x_{<t}$, with $x \in \mathcal{V}$ and alphabet size $|\mathcal{V}| = V$. For a context distribution $H \sim \mu$, this conditional distribution admits the following canonical decomposition (Huang et al., 2024)

$$p_\theta(x \mid h) = q_\theta(x)\left[1 + \sum_{k=1}^{V-1} \sigma_k^{(\theta)} \phi_k^{(\theta)}(h)\psi_k^{(\theta)}(x)\right], \tag{1}$$

where $q_\theta(x) = \mathbb{E}_{H\sim\mu}[p_\theta(x \mid H)]$ is the model's next-token marginal distribution induced over μ. The dependence on μ is suppressed in the notation.

The decomposition separates the context-independent marginal distribution q_θ from the ways context redistributes probability across tokens. The token-side canonical mode $\psi_k^{(\theta)}$ and the context-side canonical mode $\phi_k^{(\theta)}(h)$ are the paired singular functions of the ratio $p_\theta(x \mid h)/q_\theta(x) - 1$. Specifically, for the $k \le V - 1$ nonzero modes, they have zero mean and satisfy

$$\begin{aligned} \mathbb{E}_{q_\theta}[\psi_k^{(\theta)}(X)] = 0, \qquad & \mathbb{E}_\mu[\phi_k^{(\theta)}(H)] = 0, \\ \mathbb{E}_{q_\theta}[\psi_k^{(\theta)}(X)\psi_\ell^{(\theta)}(X)] = \delta_{k\ell}, \quad & \mathbb{E}_\mu[\phi_k^{(\theta)}(H)\phi_\ell^{(\theta)}(H)] = \delta_{k\ell}. \end{aligned} \tag{2}$$

Thus, the modes capture orthogonal components of the next-token prediction distribution, forming the basis of an orthogonal subspace. The singular value $\sigma_k^{(\theta)}$ quantifies the strength of the corresponding context-token dependence. Under the joint distribution induced by $H \sim \mu$ and $X \mid H \sim p_\theta(\cdot \mid H)$, it has been shown in (Huang et al., 2024),

$$\mathbb{E}[\phi_k^{(\theta)}(H)\psi_\ell^{(\theta)}(X)] = \sigma_k^{(\theta)}\delta_{k\ell}, \qquad 1 \ge \sigma_1^{(\theta)} \ge \cdots \ge \sigma_r^{(\theta)} > 0. \tag{3}$$

Therefore, the singular values are correlations between paired context and token features. Individual modes are defined up to paired sign changes and rotations within repeated singular values.

We call their ordered squares the *canonical predictive spectrum*, padding with zeros when needed:

$$\lambda_k^{(\theta)} = (\sigma_k^{(\theta)})^2, \qquad 1 \ge \lambda_1^{(\theta)} \ge \lambda_2^{(\theta)} \ge \cdots \ge 0. \tag{4}$$

[1]Throughout this paper, for brevity, we use *pretrained model* to refer to the released instruction-following foundation checkpoints used in our experiments. These models undergo large-scale next-token pretraining followed by instruction post-training.

Why probability-space geometry is distinct from logit rank. For a standard language-model head, $\ell_\theta(\cdot, h) = Wr_\theta(h) + b$, so a matrix of logits across contexts has rank at most $d + 1$, where d is the hidden dimension. Our canonical spectrum does not have this architectural rank constraint. It is defined from the realized conditional probabilities after the nonlinear softmax transformation and after removing and normalizing by the predictive marginal q_θ. A low-dimensional logit parameterization can therefore trace a nonlinear manifold in probability space whose covariance matrix has substantially higher rank. Appendix A further explains this distinction.

Information-theoretic interpretation. Let $P_\theta(h, x) = \mu(h)p_\theta(x \mid h)$. To help us further interpret this canonical predictive spectrum, we provide the following lemma, which directly connects the spectrum to information measures.

Lemma 1 *(Canonical spectrum and predictive dependence) The χ^2-information between context and the model-predicted next token induced by $P_\theta(h, x)$ satisfies*

$$\chi^2(P_\theta(h, x)\|\mu(h)q_\theta(x)) \triangleq \mathbb{E}_{H\sim\mu}\left[\sum_x \frac{(p_\theta(x \mid H) - q_\theta(x))^2}{q_\theta(x)}\right] = \sum_{k=1}^{V-1} \lambda_k^{(\theta)}. \tag{5}$$

Remark (Connection to mutual information). Shannon mutual information, which is defined as $I_\theta(H; X) \triangleq \mathbb{E}_{P_\theta}\left[\log \frac{p_\theta(X|H)}{q_\theta(X)}\right]$, also measures context-token dependence. When this dependence is weak, mutual information is approximately one-half of the χ^2-information, using natural logarithms. This approximation follows from a second-order Taylor expansion of $(1+t)\log(1+t)$ around zero.

The spectrum describes how context–token dependence is distributed across orthogonal modes: its sum equals the χ^2-information, while its concentration indicates how few modes carry most of that dependence. All eigenvalues vanish if and only if the next-token is independent of context.

2.2 Estimating the Spectrum from a Finite Dataset

Gram matrix and eigendecomposition. The population spectrum above is defined jointly by the model and a context distribution μ. It can be recovered through an eigendecomposition of the Gram matrix. To see this, define the whitened probability-deviation vector and its Gram operator

$$z_\theta(h) = D_{q_\theta}^{-1/2}\left(p_\theta(\cdot \mid h) - q_\theta\right) \in \mathbb{R}^V, \qquad C_\theta = \mathbb{E}_{h\sim\mu}\left[z_\theta(h)z_\theta(h)^\top\right]. \tag{6}$$

where $D_{q_\theta} = \mathrm{diag}(q_\theta)$. The eigendecomposition $C_\theta = U_\theta\Lambda_\theta U_\theta^\top$ recovers the canonical predictive spectrum, with eigenvalues $\lambda_k^{(\theta)} = (\sigma_k^{(\theta)})^2$. The corresponding eigenvectors give the token-side canonical modes in normalized coordinates: $u_k^{(\theta)}(x) = \sqrt{q_\theta(x)}\,\psi_k^{(\theta)}(x)$. Thus, this eigendecomposition provides both spectrum and token-side modes without computing the context-side vectors returned by a full SVD, which is extremely costly when the number of contexts is large.

Empirical plug-in estimation. In practice, we estimate these quantities from a fixed finite dataset of contexts $\mathcal{D} = \{h_i\}_{i=1}^N$, using its empirical distribution and corresponding predictive marginal:

$$\widehat{\mu}_\mathcal{D} = \frac{1}{N}\sum_{i=1}^N \delta_{h_i}, \qquad \widehat{q}_\theta(x) = \frac{1}{N}\sum_{i=1}^N p_\theta(x \mid h_i). \tag{7}$$

Here δ_{h_i} denotes a point mass at context h_i. The marginal $\widehat{q}_\theta$ averages model probabilities, not observed next-token counts. The empirical canonical spectrum is the exact spectrum of the joint law $\widehat{\mu}_\mathcal{D}(h)p_\theta(x \mid h)$. It is a plug-in estimate of the population spectrum when the dataset represents the intended context distribution. Replacing the population marginal and expectation in the Gram operator by their empirical counterparts gives

$$\widehat{z}_i = D_{\widehat{q}_\theta}^{-1/2}\big(p_\theta(\cdot \mid h_i) - \widehat{q}_\theta\big), \qquad \widehat{C}_\theta = \frac{1}{N}\sum_{i=1}^N \widehat{z}_i\widehat{z}_i^\top, \tag{8}$$

where $D_{\widehat{q}_\theta} = \mathrm{diag}(\widehat{q}_\theta)$, restricted to positive-marginal tokens. The matrix with columns $\widehat{z}_i/\sqrt{N}$ has size $V \times N$, with $N > V$. The full algorithm is presented in Algorithm 1

$$\widehat{C}_\theta = \widehat{U}_\theta\widehat{\Lambda}_\theta\widehat{U}_\theta^\top, \qquad \widehat{\Lambda}_\theta = \mathrm{diag}(\widehat{\lambda}_1^{(\theta)}, \ldots). \tag{9}$$

Algorithm 1: Canonical Predictive Geometry

Input: Next-token distributions $\{p_i = p_\theta(\cdot \mid h_i)\}_{i=1}^N$
Output: Canonical spectrum $\widehat{\Lambda}_\theta$ and token directions $\widehat{U}_\theta$
$\widehat{q}_\theta \leftarrow \frac{1}{N}\sum_{i=1}^N p_i$;
$\widehat{C}_\theta \leftarrow 0_{V\times V}$;
for $i = 1, \ldots, N$ **do**
  $\widehat{z}_i \leftarrow D_{\widehat{q}_\theta}^{-1/2}(p_i - \widehat{q}_\theta)$;
  $\widehat{C}_\theta \leftarrow \widehat{C}_\theta + \frac{1}{N}\widehat{z}_i\widehat{z}_i^\top$;
$(\widehat{U}_\theta, \widehat{\Lambda}_\theta) \leftarrow \mathrm{Eigendecom}(\widehat{C}_\theta)$;

The eigenvalues are the squared singular values, and the columns of $\widehat{U}_\theta$ are the token directions in normalized coordinates; the corresponding canonical functions are $\widehat{\psi}_k(x) = \widehat{u}_k(x)/\sqrt{\widehat{q}_\theta(x)}$.

Estimation details. The estimated geometry depends on both the model and the distribution of generated contexts. For a fixed dataset, each model generates its own autoregressive trajectories and therefore induces its own empirical context distribution and predictive marginal $\widehat{q}_\theta$. Consequently, both the canonical eigenvalues and token directions can vary with the model and the probing dataset.

We estimate each geometry from $N = 10^6$ generated contexts. Despite eliminating the context dimension before eigendecomposition, Algorithm 1 still requires decomposing a $V \times V$ token-side covariance matrix, with $O(V^2)$ memory complexity. For modern LLM vocabularies, this matrix can be enormous (e.g., $150k \times 150k$), making standard single-node dense eigendecomposition computationally impractical. To overcome this challenge, we develop a distributed algorithm that computes the full symmetric eigendecomposition in double precision using ScaLAPACK `PDSYEVD`. The covariance matrix is distributed across MPI processes using a two-dimensional block-cyclic layout.

3 Foundation LLMs Exhibit a Low-Dimensional Predictive Geometry

The proposed framework in Section 2 turns the dimensionality of next-token prediction of LLMs into a spectral analysis problem, allowing us to study the question: *how many directions in token space does an LLM actually use as its predictions change with context?* A priori, there is little reason to expect the answer to be small. Modern language models (Liu et al., 2024; Touvron et al., 2023; Bai et al., 2023) predict over vocabularies containing tens or even hundreds of thousands of tokens, and context could in principle reshape the next-token distribution along a correspondingly large collection of independent directions. In the canonical spectrum, this would appear as substantial eigenvalues extending throughout a broad portion of the available token space.

3.1 Experimental Setup

Models and datasets. We study Meta-Llama-3.1-8B-Instruct and Qwen2.5-14B-Instruct, whose large vocabularies provide high-dimensional ambient token spaces in which to study predictive geometry. Our main spectral analysis uses GSM8K and MMLU, providing two substantially different distributions of problems and generated reasoning trajectories. We additionally use ELI5 and GSM8K in the post-training experiments described later in this section.

Sampling predictive contexts. For each model–dataset configuration, we construct the context distribution through autoregressive generation. Given a dataset example x, we append the instruction *"think step by step"* and let the model generate up to 512 tokens. Throughout our experiments, we generate trajectories using deterministic greedy decoding ($\tau = 0$), without top-p or top-k truncation.

At generation step t, the context

$$h_t = (x, \text{“think step by step”}, x_1, \ldots, x_{t-1})$$

contains both the original input and the model's previously generated tokens. We evaluate and store the model's full next-token probability distribution $p_\theta(\cdot \mid h_t) \in \mathbb{R}^V$ before generating the next token. We then append the generated token to the context and repeat this process. After completing each trajectory, we move to the next dataset example until we collect $N = 10^6$ next-token contexts.

Importantly, the resulting context distributions are model-dependent. Although models are evaluated using the same dataset and prompting protocol, their generated trajectories may differ, and therefore each model induces its own empirical distribution over contexts.

Estimating the canonical geometry. From the resulting 10^6 next-token distributions, we estimate the model-specific predictive marginal $\widehat{q}_\theta$, construct the whitened probability deviations $\widehat{z}_i$ and form $\widehat{C}_\theta$. Then, we eigendecompose $\widehat{C}_\theta$ as described in Section 2.2. Its ordered eigenvalues define the canonical predictive spectrum, while the corresponding eigenvectors define the token-side predictive geometry used in the subspace analyses below.

3.2 Next-Token Prediction Is Low-Dimensional

We first ask how many token-space directions carry substantial context-dependent predictive variation in a foundation LLM. If next-token prediction uses a large portion of the available vocabulary space, appreciable eigenvalues would span a large fraction of the V canonical directions. Conversely, concentration of spectral energy in the leading modes indicates a low-rank predictive geometry.

We estimate the canonical spectrum independently for Meta-Llama-3.1-8B-Instruct and Qwen2.5-14B-Instruct (Team, 2024; Yang et al., 2024) using the GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021) generation distributions described in Section 3.1. For each model–dataset pair, the geometry is estimated from 10^6 autoregressive contexts. Figure 1 shows the resulting ordered eigenvalues. Across both models and datasets, the spectrum is strongly concentrated: the leading canonical modes carry most of the context-dependent predictive variation, followed by a broad spectral tail with negligible eigenvalues.

Effective predictive dimension. To quantify this concentration, we define the cumulative predictive energy

$$\mathcal{E}_\theta(r) = \frac{\sum_{k=1}^{r} \widehat{\lambda}_k^{(\theta)}}{\sum_{k=1}^{V} \widehat{\lambda}_k^{(\theta)}},$$

and the effective dimension at level α,

$$r_\alpha = \min\{r : \mathcal{E}_\theta(r) \geq \alpha\}.$$

We use $r_{0.999}$ to denote the number of canonical modes required to capture 99.9% of the total context-dependent predictive variation. Table 1 quantifies the spectral concentration shown in Figure 1. On GSM8K, only 13,476 modes for Llama and 14,716 modes for Qwen are needed to capture 99.9% of predictive variation, representing just 10.51% and 9.68% of their respective vocabulary dimensions. On MMLU, the effective dimensions are higher, requiring 30,787 modes for Llama and 32,142 for Qwen, or 24.00% and 21.14% of their vocabulary dimensions, respectively. Across both models and datasets, even at the stringent 99.9% threshold, most token-space directions contribute negligibly to context-dependent predictive variation.

The effective dimension depends on the context distribution: MMLU consistently requires a larger fraction of token space than GSM8K for both model families. This suggests that different tasks induce different levels of complexity in next-token prediction, with MMLU exhibiting more diversity in context-dependent predictive variation. Nevertheless, despite these differences in effective dimension, both model families exhibit strong spectral concentration across datasets. This suggests that the low-dimensional organization of next-token prediction is a consistent property across the models and context distributions studied, rather than a phenomenon specific to a particular task.

3.3 The Low-Dimensional Geometry Is Learned

The preceding result establishes that foundation LLMs exhibit a low-dimensional predictive geometry, but leaves open how this structure emerges. In particular, language-model logits are already

Table 1: **Effective dimension of next-token prediction.** $r_{0.999}$ denotes the smallest number of canonical modes accounting for 99.9% of the total spectral energy. Additional threshold analyses are presented in Appendix C

Model	Dataset	$r_{0.999}$	$r_{0.999}/V$
Llama-3.1	GSM8K	13,476	10.51%
Llama-3.1	MMLU	30,787	24.00%
Qwen2.5	GSM8K	14,716	9.68%
Qwen2.5	MMLU	32,142	21.14%

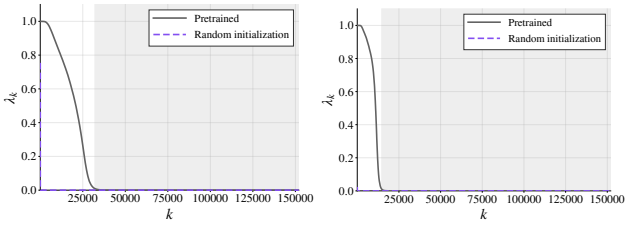


Figure 2: **Training induces a structured predictive spectrum.** Ordered canonical eigenvalues for Qwen2.5-14B-Instruct on MMLU (left) and GSM8K (right). Solid curves denote foundation models and dashed curves denote randomly initialized models with the same architecture. The shaded region denotes the spectral tail beyond $r_{0.999}$.

subject to dimensional constraints induced by the hidden representation and linear output map (Yang et al., 2017; Godey & Artzi, 2026; Godey et al., 2024). We therefore ask whether the observed canonical structure is simply an architectural consequence or instead emerges through large-scale pre-training.

To isolate this effect with our framework, we compare each foundation model with a randomly initialized model having the same architecture, tokenizer, vocabulary, and output parameterization. Random and trained models are evaluated on the same dataset and with the same prompting and autoregressive generation protocol.

Figure 2 and results in Appendix B show that the two regimes are qualitatively different. Randomly initialized models exhibit little context-dependent predictive structure beyond a small number of leading components. After pre-training, a much larger set of canonical modes becomes active, but this variation remains confined to a restricted portion of token space.

The observed low-dimensional geometry is therefore not explained by the architecture alone. Pre-training creates substantial context-dependent predictive variation while organizing that variation within a limited set of token-side directions. This is distinct from the usual low-rank constraint on logits: our spectrum is computed from the model's realized next-token probability distributions after the nonlinear softmax transformation and relative to the model-induced predictive marginal.

3.4 Fine-Tuning Preserves the Dominant Predictive Subspace

Beyond characterizing the predictive geometry of pre-trained foundation LLMs, our framework can be used to study how this geometry changes under downstream adaptation. In this subsection, we ask two questions: does fine-tuning preserve the low-dimensional organization of predictive variation, and, more strongly, does it preserve the dominant token-space directions themselves? We fine-tune Qwen2.5-14B-Instruct independently on GSM8K and ELI5 and re-estimate its canonical geometry before and after adaptation. Each geometry is estimated from 10^6 autoregressive contexts generated using the protocol in Section 3.1. Because generation is model-dependent, the resulting empirical context distribution can differ before and after fine-tuning.

Fine-tuning setup. We fine-tune Qwen2.5-14B-Instruct separately on GSM8K and ELI5 using LoRA-based supervised fine-tuning. LoRA adapters are applied to all attention and MLP projection layers with rank 64 and scaling factor 128. Both models are trained for 100 optimization steps with learning rate 5×10^{-6}, effective batch size 16, and maximum sequence length $16{,}000$; all reported spectral analyses use the checkpoint at step 100. For GSM8K, we train on the official training split using the original step-by-step rationales as targets, with final answers formatted using `\boxed{}`. For ELI5, we use the `sentence-transformers/eli5` question–answer pairs, with a 90/10 train–validation split (seed 42), and train on the provided long-form answers.

Does fine-tuning preserve the low-dimensional spectrum? We first compare the canonical eigenvalue spectra before and after fine-tuning. If adaptation introduces substantially new predictive directions, spectral energy should spread into modes that were weak in the foundation model.

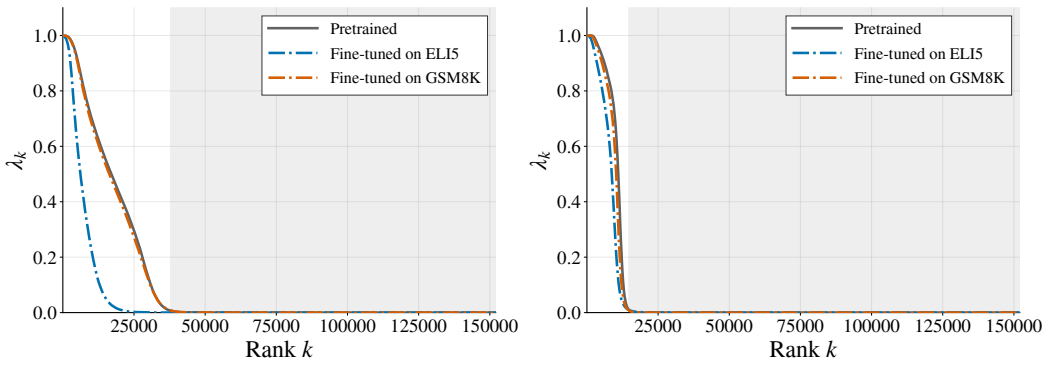


Figure 3: **Fine-tuning preserves the low-dimensional canonical spectrum.** Canonical eigenvalue spectra of Qwen2.5-14B-Instruct before and after fine-tuning, estimated using generations from ELI5 (left) and GSM8K (right). The shaded region starts at the minimum rank required to capture 99.9% of the foundation model's total spectral energy.

Figure 3 shows that fine-tuning changes the magnitudes of individual modes but leaves the overall spectral concentration largely unchanged. In both evaluation distributions, most predictive variation remains concentrated in the same restricted range of leading modes. Fine-tuning therefore preserves the low-dimensional organization of next-token prediction. More interestingly, the fine-tuned spectra decay slightly earlier, suggesting that supervised fine-tuning further concentrates predictive variation rather than expanding it into previously inactive directions. Thus, adaptation appears to specialize the model largely within the predictive geometry already established by the foundation model, while suppressing some of its weaker modes.

Does fine-tuning preserve the same canonical token subspace? Similar spectra do not imply similar geometry: two models can concentrate predictive variation in the same number of dimensions while using different token-space directions. We therefore compare the leading canonical subspaces directly.

For a probing dataset D, let $U^D_{\text{base},r}$ denote the leading r canonical directions of the foundation model, and let $U^D_{\text{ft},r}$ denote the corresponding directions after fine-tuning. Each model's canonical operator is constructed using its own predictive marginal $\widehat{q}_\theta$, as defined in Section 2.2. Because all comparisons use the same tokenizer and vocabulary, the resulting eigenvectors share the same token-index coordinate system. The comparison therefore measures alignment between the models' marginal-normalized canonical predictive subspaces.

We quantify this alignment using

$$\text{Overlap}_r(U_A, U_B) = \frac{1}{r} \left\| U_A^\top U_B \right\|_F^2 . \tag{10}$$

This quantity equals one when the two subspaces coincide and zero when they are orthogonal. For independent isotropic r-dimensional subspaces in $\mathbb{R}^V$, its expected value is r/V.

Figure 4 shows substantial alignment between the foundation and fine-tuned canonical subspaces throughout the active part of the spectrum. For Qwen2.5-14B-Instruct, 99.9% of the foundation model's spectral energy is contained within approximately $r \approx 32$k modes. At this rank, the fine-tuned models retain roughly 0.7–0.9 overlap with the foundation-model subspace, compared with $r/V \approx 0.21$ for unrelated isotropic subspaces in a vocabulary of approximately 152k tokens.

The preservation also extends across tasks. When the foundation-model canonical geometry is estimated from ELI5 generations, substantial alignment remains after GSM8K fine-tuning; similarly, a geometry estimated from GSM8K generations remains strongly aligned after ELI5 fine-tuning. Thus, even when the downstream task changes, fine-tuning retains a substantial fraction of the dominant marginal-normalized predictive geometry of the foundation model.

Interestingly, the spectral boundary $r_{0.999}$, defined solely from the eigenvalues of the foundation model, also marks a clear transition in the subspace-overlap curves. Within the active spectral region

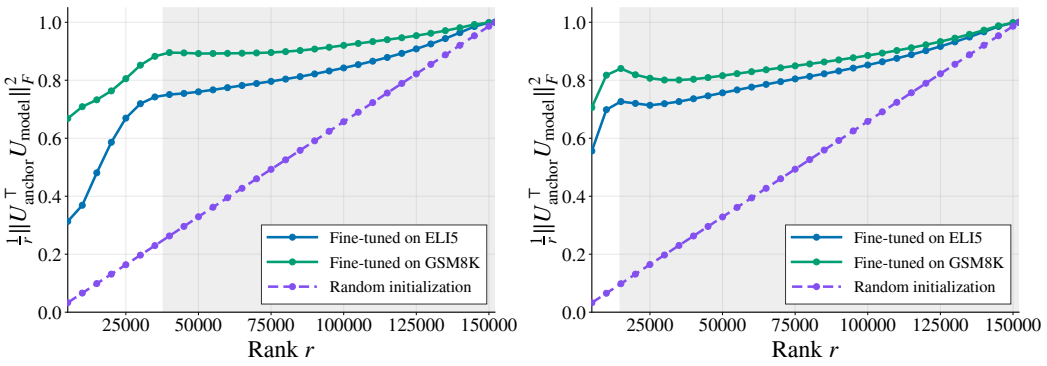


Figure 4: **Fine-tuning preserves much of the dominant token subspace.** Overlap with the foundation-model canonical subspace estimated on ELI5 (left) and GSM8K (right). Curves compare the anchor with models fine-tuned on ELI5 and GSM8K and with a randomly initialized model. The shaded region begins at the rank containing 99.9% of the foundation model's cumulative spectral energy.

($r < r_{0.999}$), overlap rises rapidly as the leading predictive directions are added. Around $r_{0.999}$, the curves enter a second regime in which additional directions yield only gradual changes in overlap. Thus, the same rank that captures essentially all predictive energy also separates a strongly preserved canonical subspace from a much weaker spectral tail. This agreement between the eigenvalue and eigenvector analyses suggests that $r_{0.999}$ identifies not only an effective predictive dimension, but also a natural boundary of the geometrically stable subspace preserved under fine-tuning.

Together, the spectral and subspace analyses establish two complementary forms of preservation. Fine-tuning retains the low-dimensional concentration of context-dependent predictive variation and, more strongly, much of the canonical token subspace carrying that variation. These results motivate the **Canonical Subspace Hypothesis**: *language-model pretraining organizes next-token prediction around a low-dimensional canonical token geometry that downstream adaptation largely reuses rather than reconstructs*.

4 Conclusion

In this paper, we introduced a spectral framework for characterizing the geometry of LLM next-token prediction directly in probability space. By separating context-dependent variation from the model's predictive marginal, the framework decomposes next-token distributions into orthogonal canonical token modes and provides complementary measures of their spectral concentration and geometric orientation. Using this framework, we find that foundation LLMs exhibit a pronounced low-dimensional predictive geometry: despite operating over vocabularies with more than 10^5 tokens, most context-dependent predictive variation is concentrated in a restricted set of dominant token-space directions. Across both model families and multiple context distributions, the canonical spectrum remains strongly concentrated in this restricted region.

Beyond measuring the effective dimensionality of next-token prediction, our framework allows us to directly compare the predictive geometry of a foundation LLM and its fine-tuned variants. We find that fine-tuning changes $p_\theta(\cdot \mid h)$ while largely preserving both the concentration of the canonical spectrum and the dominant token-side canonical subspace, including under cross-task adaptation. $r_{0.999}$ also marks a geometric transition. These results motivate the **Canonical Subspace Hypothesis**: *language-model pre-training organizes next-token prediction around a low-dimensional token geometry that downstream adaptation largely reuses rather than reconstructs*. By making this subspace explicit, our framework provides a geometric view of efficient LLM adaptation: fine-tuning can modify the model's conditional next-token distributions and downstream behavior while retaining much of the predictive basis learned by the foundation model. The framework is general and applies directly to the next-token distributions produced by any LLM. Future work can use it, for example, to compare how different post-training procedures, such as supervised fine-tuning and GRPO, reshape $p_\theta(\cdot \mid h)$, or to study geometric scaling laws by tracking how the canonical spectrum and dominant token-side subspace evolve with model scale.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and presentation of the manuscript. All AI-assisted manuscript polishing was manually reviewed and revised for technical correctness and consistency with the reported results. The authors take full responsibility for the final content of the paper.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, pp. 7319–7328, 2021.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Basile Confavreux, Aaditya Singh, Jin Hwa Lee, Amaury Sabran, and Andrew Saxe. Comparing the learning dynamics of in-context learning and fine-tuning in language models. In *International Conference on Learning Representations*, volume 2026, pp. 157425–157437, 2026.

Nathan Godey and Yoav Artzi. Lost in backpropagation: The lm head is a gradient bottleneck. *arXiv preprint arXiv:2603.10145*, 2026.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Why do small language models underperform? studying language model saturation via the softmax bottleneck. *arXiv preprint arXiv:2404.07647*, 2024.

Noah Golowich, Allen Liu, and Abhishek Shetty. Sequences of logits reveal the low rank structure of language models. In *International Conference on Learning Representations*, volume 2026, pp. 25335–25371, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Shao-Lun Huang, Anuran Makur, Gregory W Wornell, and Lizhong Zheng. Universal features for high-dimensional learning and inference. *Foundations and Trends® in Communications and Information Theory*, 21(1-2):1–299, 2024.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, volume 2024, pp. 9944–10015, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Liam Madden, Curtis Fox, and Christos Thrampoulidis. Next-token prediction capacity: general upper bounds and a lower bound for transformers. *IEEE Transactions on Information Theory*, 71 (9):7134–7148, 2025.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, et al. Less is more: Local intrinsic dimensions of contextual language models. *Advances in Neural Information Processing Systems*, 38:67795–67828, 2026.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Haojin Wang, Zining Zhu, and Freda Shi. Distribution prompting: Understanding the expressivity of language models through the next-token distributions they can produce. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20915–20928, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*, 2017.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

A WHY PROBABILITY-SPACE GEOMETRY RATHER THAN LOGIT RANK?

A natural alternative to our canonical analysis is to study low-rank structure directly in the model's logits. Here we clarify why the two notions characterize different geometric objects.

Architectural rank constraint in logit space. Let $\ell_\theta(h) \in \mathbb{R}^V$ denote the vector of pre-softmax logits produced at context h. For a standard language-model head,

$$\ell_\theta(h) = W r_\theta(h) + b, \tag{11}$$

where $r_\theta(h) \in \mathbb{R}^d$ is the final hidden representation, $W \in \mathbb{R}^{V\times d}$, and $b \in \mathbb{R}^V$. For a collection of N contexts, let

$$L_\theta = [\ell_\theta(h_1) \quad \cdots \quad \ell_\theta(h_N)] \in \mathbb{R}^{V\times N}.$$

Then

$$L_\theta = W R_\theta + b\mathbf{1}^\top, \qquad \operatorname{rank}(L_\theta) \le d+1, \tag{12}$$

where $R_\theta = [r_\theta(h_1), \ldots, r_\theta(h_N)]$. Thus, low rank of the ordinary next-token logit matrix is directly constrained by the architecture through the hidden dimension d, closely related to the softmax bottleneck (Yang et al., 2017).

Our canonical spectrum measures a different object. It is constructed from the realized conditional probabilities

$$p_\theta(\cdot \mid h) = \operatorname{softmax}(\ell_\theta(h)) \tag{13}$$

after separating the model's predictive marginal $q_\theta = \mathbb{E}_{h\sim\mu}[p_\theta(\cdot \mid h)]$. The softmax map is nonlinear. Consequently, the architectural rank bound in equation 12 does not impose an analogous $d+1$ bound on the linear span or covariance of the resulting probability vectors.

Importantly, this does not mean that softmax increases the intrinsic manifold dimension of the model outputs. The image of a d-dimensional logit parameterization remains at most d-dimensional as a nonlinear manifold. However, a low-dimensional curved manifold can have a much higher-dimensional linear span and covariance. Our canonical spectrum characterizes this latter linear organization of context-dependent probability variation, rather than the intrinsic dimension of the underlying nonlinear manifold.

Canonical geometry of probability deviations. Recall that our analysis uses the marginal-normalized probability deviation

$$z_\theta(h) = D_{q_\theta}^{-1/2}\big(p_\theta(\cdot \mid h) - q_\theta\big), \tag{14}$$

and the corresponding token-side covariance

$$C_\theta = \mathbb{E}_{h\sim\mu}\left[z_\theta(h) z_\theta(h)^\top\right]. \tag{15}$$

For a finite set of contexts, the matrix whose columns are $z_\theta(h_i)/\sqrt{N}$ can be written as

$$B_\theta = \left[\tfrac{z_\theta(h_1)}{\sqrt{N}} \quad \cdots \quad \tfrac{z_\theta(h_N)}{\sqrt{N}}\right], \qquad C_\theta = B_\theta B_\theta^\top. \tag{16}$$

Its entries are

$$B_\theta(x,i) = \frac{1}{\sqrt{N}} \frac{p_\theta(x \mid h_i) - q_\theta(x)}{\sqrt{q_\theta(x)}}. \tag{17}$$

Thus, the canonical eigenspectrum is determined by fluctuations of the *normalized probabilities* around their predictive marginal, rather than by linear dependence among the pre-softmax scores.

Low rank of L_θ therefore does not determine the spectral concentration of C_θ. Conversely, concentration of the canonical spectrum does not imply an additional low-rank constraint on the logits. The two analyses concern different spaces and answer different questions: logit rank characterizes the linear parameterization of pre-softmax scores, whereas the canonical spectrum characterizes how the model's realized probability distributions vary with context.

Relation to log-likelihood-ratio geometry. A stronger alternative might be to characterize context dependence in log-probability space rather than through raw logits. We therefore ask whether low-dimensional structure in the log-likelihood ratio $\log p_\theta(x \mid h)/q_\theta(x)$ would already imply the low-dimensional canonical geometry studied here. The distinction can also be expressed through the log-likelihood ratio

$$s_\theta(h, x) = \log \frac{p_\theta(x \mid h)}{q_\theta(x)}. \tag{18}$$

A low-dimensional model in natural-parameter space may, for example, take the form

$$p_\theta(x \mid h) = q_\theta(x) \exp\left(\sum_{k=1}^{r} \tau_k f_k(h) g_k(x) - A_\theta(h)\right), \tag{19}$$

where $A_\theta(h)$ is the context-dependent normalizer. Equivalently,

$$s_\theta(h, x) = \sum_{k=1}^{r} \tau_k f_k(h) g_k(x) - A_\theta(h). \tag{20}$$

This describes low-dimensional structure in log-probability, or natural-parameter, space.

By contrast, the canonical decomposition is linear in the *relative probability deviation*:

$$\frac{p_\theta(x \mid h)}{q_\theta(x)} - 1 = \sum_k \sigma_k^{(\theta)} \phi_k^{(\theta)}(h) \psi_k^{(\theta)}(x). \tag{21}$$

The two representations are connected exactly by

$$\frac{p_\theta(x \mid h)}{q_\theta(x)} - 1 = \exp(s_\theta(h, x)) - 1. \tag{22}$$

Indeed, even a rank-one log-likelihood ratio $s(h, x) = f(h)g(x)$ generates the sequence of components $f(h)^m g(x)^m$ through the expansion of $e^s - 1$, and can therefore produce much higher rank in probability space. Because elementwise exponentiation is nonlinear, it does not preserve matrix rank or singular directions. Hence, low-rank structure in $s_\theta(h, x)$ does not in general imply the same low-rank structure in the relative probability deviations appearing in equation 21.

The two quantities agree locally. When $|s_\theta(h, x)|$ is small,

$$\exp(s_\theta(h, x)) - 1 = s_\theta(h, x) + O\big(s_\theta(h, x)^2\big). \tag{23}$$

Thus, in the weak-dependence regime, the relative probability deviation and the log-likelihood ratio agree to first order. Away from this regime, however, their geometries need not coincide.

What the canonical spectrum measures. These distinctions clarify the object studied in this work. The canonical geometry is invariant to context-dependent additive shifts of the logits, since such shifts leave $p_\theta(\cdot \mid h)$ unchanged. It also separates the context-independent predictive marginal q_θ from the ways in which context redistributes probability across tokens. Moreover, unlike the architectural rank bound on logits, the canonical spectrum admits the exact dependence interpretation

$$\sum_k \lambda_k^{(\theta)} = \chi^2(P_\theta(H, X) \,\|\, \mu(H) q_\theta(X)), \tag{24}$$

established in Lemma 1.

Therefore, the low-dimensional phenomenon studied in this paper is not the statement that LLM logits are low rank. Rather, it is the empirical finding that the *context-dependent variation of the realized next-token probabilities*, after accounting for the predictive marginal, is concentrated in a restricted set of canonical token directions. Recent analyses of low-dimensional structure in sequential log-probabilities study a complementary logit- or natural-parameter-space phenomenon (Golowich et al., 2026); our focus is specifically the geometry of the next-token predictive distribution itself.

B RANDOM VS PRETRAINED ANALYSIS

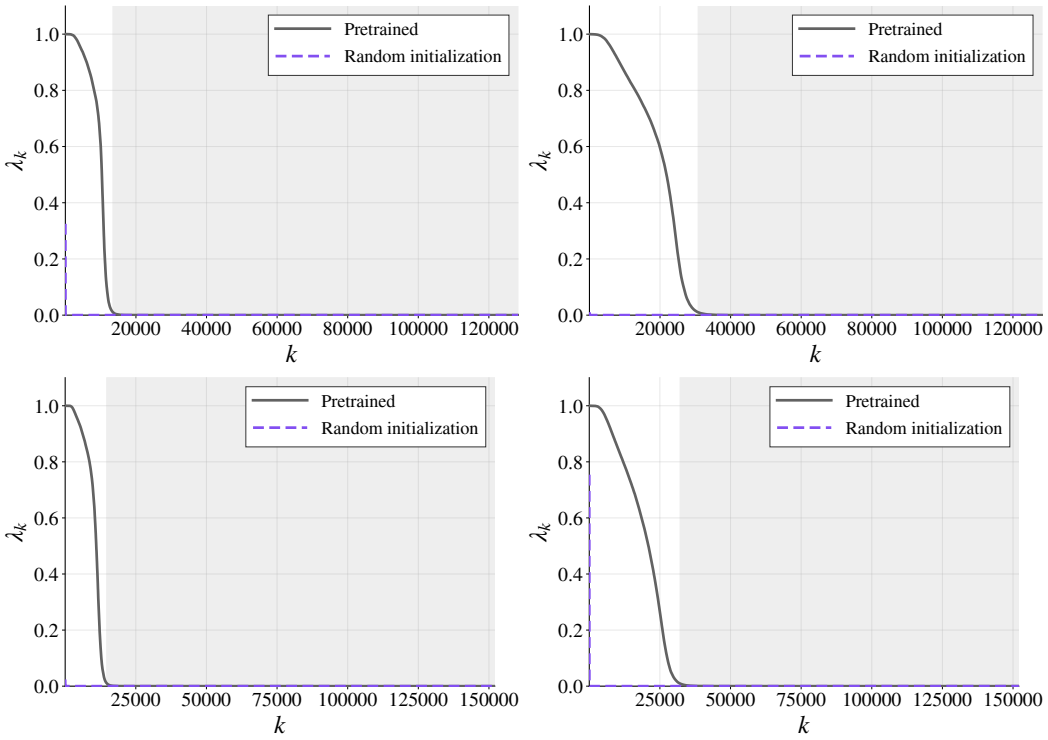


Figure 5: **Training induces a structured predictive spectrum.** Ordered canonical eigenvalues for Meta-Llama-3.1-8B-Instruct (up) and Qwen2.5-14B-Instruct (down) on GSM8K (left) and MMLU. Solid curves denote foundation models and dashed curves randomly initialized models with the same architecture. Random initialization produces an almost degenerate spectrum, whereas trained models develop a broad set of active canonical modes while remaining strongly concentrated relative to the full vocabulary dimension. The shaded region denotes the spectral tail beyond the rank containing 99.9% of the foundation model's cumulative spectral energy.

C Analysis of Different Thresholds for Effective Canonical Dimension

Table 2: **Effective canonical dimension at different spectral energy thresholds.** For each model and dataset, r_α denotes the smallest number of canonical modes required to capture an α fraction of the total spectral energy. Percentages indicate the fraction of the vocabulary dimension V.

Model	Dataset	V	$r_{0.80}$	$r_{0.90}$	$r_{0.95}$	$r_{0.99}$
Llama-3.1-8B	GSM8K	128,256	8,212 (6.40%)	9,495 (7.40%)	10,270 (8.01%)	11,553 (9.01%)
Llama-3.1-8B	MMLU	128,256	17,942 (13.99%)	21,160 (16.50%)	23,240 (18.12%)	26,536 (20.69%)
Qwen2.5-14B	GSM8K	152,064	8,830 (5.81%)	10,245 (6.74%)	11,158 (7.34%)	12,636 (8.31%)
Qwen2.5-14B	MMLU	152,064	18,027 (11.85%)	21,565 (14.18%)	23,960 (15.76%)	27,631 (18.17%)