Spectral Signatures of Large Language Models

Zhuoying Zhang [*1] **Ishan Verma Prasad** [*2] **Yuanzhe Hu** [3] **Zihang Liu** [4] **Hengrui Luo** [1] **Pu Ren** [5]
Yaoqing Yang [2]

Abstract

The rapidly growing repository of publicly available large language models (LLMs) presents significant challenges for systematic management and quantification at scale, such as model lineage tracing, licensing, and evaluation. However, task-specific benchmarks are insufficient for this setting, as LLMs differ widely in architectures, scales, and training procedures. To address this challenge, we adopt *spectral shape*-based metrics for managing and quantifying LLMs based on Heavy-Tailed Self-Regularization theory. Our approach uses the shape information of the weight empirical spectral density as a compact spectral signature of each model. This signature captures intrinsic properties of pretrained models and remains robust during post-training, making it suitable for model-level analysis. In addition, this metric is *data-free*, *computationally-efficient*, and *scale-invariant*, enabling large-scale analysis in practice. Moreover, we curate a large and diverse model corpus consisting of major open-source LLM families, and use it to systematically benchmark spectral and non-spectral metrics across models and downstream tasks. We show that our spectral signature supports the tracking of the model lineage, the unsupervised clustering of similar models, and the quantification of the model performance. Overall, the proposed spectral signature provides a meaningful proxy for broad performance trends across LLMs, enabling efficient organization, comparison, and analysis of large model collections. The code is available at `https://github.com/Ingrid-505/Spectral_Signature`.

[1]Rice University, Houston, TX, USA [2]Dartmouth College, Hanover, NH, USA [3]Georgia Institute of Technology, Atlanta, GA, USA [4]University of California Berkeley, Berkeley, CA, USA [5]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. Correspondence to: Pu Ren <pren@lbl.gov>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Publicly available large language models (LLMs) have been growing rapidly in recent years, with hundreds of thousands of models now available through open-source platforms such as Hugging Face (HF) (Zhuang et al., 2024). These models exhibit significant diversity in their training trajectories, including variations in model architecture, parameter scale, training data, and empirical performance (Horwitz et al., 2025a). As the ecosystem grows, there is an increasing need for reliable and efficient methods to characterize, compare, and manage LLMs beyond task-level benchmarks. Representative examples include identifying whether deployed LLM systems can be traced to the same backbone foundation model for intellectual property and safety purposes (Zhu et al., 2025), as well as tracking and organizing model variants and versions in large-scale LLM databases (Horwitz et al., 2025b). Moreover, LLMs present distinct idiosyncrasies across downstream tasks, with different models specializing in different areas such as code generation, mathematical reasoning, long-context processing, or interactive dialogue. These considerations point to the need for a *principled* and *data-light* framework to trace model sources, capture model idiosyncrasies, and quantify model performances for large-scale LLM collections.

In this work, we introduce a *spectral shape*-based perspective for LLM quantification that builds on insights from Heavy-Tailed Self-Regularization (HT-SR) theory (Martin & Mahoney, 2021; Martin et al., 2021; Martin & Hinrichs, 2025). HT-SR theory analyzes the eigenspectrum of weight matrices, and prior studies (Yang et al., 2023; Martin et al., 2021) have shown that the shape of the empirical spectral density (ESD) encodes informative signals about model capacity, quality, and training dynamics that are not well captured by scale-based measures. Moreover, ESD shape metrics are data-independent, computationally lightweight, and remain robust under noise and scaling. These properties make them particularly well-suited to large-scale LLM settings, where task-specific evaluation can be expensive. An overview of different quantification metrics is shown in Tab. 1.

We use the shape metric of the ESD, namely PL_Alpha_Hill, as a *spectral signature* of each LLM. The PL_Alpha_

Hill metric summarizes the global structure of layer-wise weight eigenspectra into a compact representation, enabling efficient storage and comparison across large model collections. We adopt this spectral signature to support three downstream tasks: supervised model classification, unsupervised clustering of LLMs, and quantifying LLM performance. Our contributions are summarized below.

- We introduce a shape-based metric for reliably and efficiently characterizing, comparing, and quantifying LLMs beyond task-level benchmarks. Our metric is *scale-invariant*, *data-free*, *theory-driven*, and *easy-to-compute*.
- We curate a diverse and representative LLM corpus of up to 499 models and use it to benchmark three tasks: LLM similarity measurement, model classification and clustering, and performance prediction. We further conduct a systematic comparison of our spectral signature against a range of baselines across various experimental settings, demonstrating *consistent effectiveness* across tasks and conditions.
- We compare shape- and scale-based metrics under output-invariant reparameterizations and noise perturbations. ESD shape metrics remain stable, while scale-based baselines degrade, indicating that spectral shape captures intrinsic weight structure for reliable LLM quantification.

2. Related Work

Model Lineage and Independence Testing. The growing population of open-source LLMs has spurred research into determining LLM lineage—whether two models are independently trained or one is trained from another—which is crucial for intellectual property and model safety. Recent works use distinct approaches: output-based methods use output comparison (Nikolic et al., 2025), fingerprint queries (Jin et al., 2024), and lexical cues (Sun et al., 2025; Wu et al., 2023) to compare models without requiring parameter access; watermarking techniques inject provenance information during training or decoding (Wang & Kerschbaum, 2019; Uchida et al., 2017; Gu et al., 2022; Jiang et al., 2025); whereas white-box methods perform hypothesis testing (Zhu et al., 2025) or fingerprinting using metrics derived from model weight structures (Zeng et al., 2024; 2025; Wang et al., 2025; Yoon et al., 2025).

Model Zoos and LLM Embeddings. There has been growing interest in exploring representation-based approaches to capture model dependence, properties, and relationships, enabling efficient organization and management of large model zoos. Relevant directions include LLM embeddings (Zhuang et al., 2024), weight-space learning (Horwitz et al., 2025a;b; Gueta et al., 2023), and loss-landscape-based grouping (Gueta et al., 2023; Schürholt et al., 2025). LLM embedding methods (Zhuang et al., 2024) learn compact and model-level representations that can be used to predict downstream performance across tasks without additional inference. In contrast, weight-space learning approaches operate directly on model parameters. For example, Horwitz *et al.* (Horwitz et al., 2025a) combine weight-space distances with model metadata to construct an "atlas" of open models, enabling similarity search and provenance analysis within large repositories. A complementary line of work (Schürholt et al., 2025) considers loss-landscape and phase-based information (Yang et al., 2021), such as mode connectivity and CKA similarity, to partition models into distinct regimes and assess their similarity.

HT-SR Theory. HT-SR theory is a theoretical framework based on random matrix theory for analyzing weight matrices geometry of neural networks (NNs) (Mahoney & Martin, 2019; Martin & Mahoney, 2021; Couillet & Liao, 2022; Liao & Mahoney, 2025). HT-SR theory was proposed based on the observation that well-trained, state-of-the-art NNs often exhibit HT structures in the ESD of each layer. In the meantime, several rigorous theories in stochastic gradient descent relating HT phenomena to generalization performance were established, providing further theoretical support for HT-SR theory (Gurbuzbalaban et al., 2021; Hodgkinson & Mahoney, 2021; Hodgkinson et al., 2022; Simsekli et al., 2019; 2020). Recent works extend theoretical foundations and show that HT-SR Theory can be used to assess the training quality of modern deep NN models across various domains (e.g., computer vision and natural language processing) without accessing any training or test data (Yang et al., 2023; Martin & Mahoney, 2021). Based on this insight, recent studies have further proposed HT-SR theory-based methods for practical NN applications, including hyperparameter tuning (Zhou et al., 2023; Liu et al., 2024; Hu et al., 2025) and model pruning (Lu et al., 2024; Hu et al., 2025). In this work, we use the HT-SR theory to analyze the geometric similarities of weight matrices of LLMs.

3. Preliminaries

Problem Setup. In large LLM repositories, many checkpoints share a common pretrained backbone but differ through post-training, distillation, pruning/quantization, or model merging. Our goal is to perform *data-free* lineage analysis: comparing models and tracing backbone relationships directly from their weights. This requires robustness to output-invariant transformations, such as uniform scaling,

Table 1. Comparisons of different metrics for LLM tasks. ✓ and ✗ denote strong and weak performance, respectively.

Metrics	Scale-invariant	Data-free	Easy to compute	Scalable	Theory-driven	Classification	Clustering	Quantification
Logits (Yang & Wu, 2024)	✓	✗	✓	✗	✗	✓	✓	✗
REEF (Zhang et al., 2025)	✓	✗	✓	✗	✗	✓	✓	✗
PCS (Zeng et al., 2024)	✗	✓	✓	✗	✗	✓	✓	✗
Spectral norm	✗	✓	✓	✓	✓	✓	✓	✓
Log alpha norm	✗	✓	✓	✓	✓	✓	✓	✓
Shape metric (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

hidden-unit permutations, and linear-chain reparameterizations, which can make naive weight-space similarities unreliable. We therefore focus on spectral shape statistics of weight matrices, which capture intrinsic weight structures while being less sensitive to transformations that change the parameterization but preserve the model's input-output function.

Weight Analysis with HT-SR Theory. HT-SR theory (Martin & Mahoney, 2021) demonstrates the empirical fact that well-trained NNs tend to exhibit strong correlations in weight matrices, resulting in a heavy-tailed ESD of each weight matrix. ESD characterizes the eigenvalue distribution of the correlation matrix $\mathbf{W}^\top\mathbf{W}$ of the weight matrix $\mathbf{W}$. HT-SR theory can be used to analyze the dynamics of weight matrix ESDs during training, by modeling the tail component of the ESD with a power-law (PL) distribution within the interval $(\lambda_{min}, \lambda_{max})$, and its tail index α: $p(\lambda) \propto \lambda^{-\alpha}, \quad \lambda_{min} < \lambda < \lambda_{max}$. One commonly used method to approximate the tail index α is with the Hill Estimator[1] (Hill, 1975), which computes an estimate PL_Alpha_Hill:

$$\text{PL_Alpha_Hill} = 1 + \frac{k}{\sum_{i=1}^{k} \ln \frac{\lambda_{n-i+1}}{\lambda_{n-k}}}, \tag{1}$$

where $\{\lambda_i\}_{i=1}^{n}$ is sorted in ascending order. For the selection of k, we adopt the approach from previous works (Zhou et al., 2023; Yang et al., 2023), where we set $k = \frac{n}{2}$. PL_Alpha_Hill measures the shape of the eigenspectrum of each layer individually. Unlike norm-based metrics such as the spectral norm or RMS norm, PL_Alpha_Hill is robust under scaling. Moreover, since permutations are orthogonal changes of basis, the spectrum of $\mathbf{W}^\top\mathbf{W}$ (and hence its ESD shape) is invariant to hidden-unit permutations, motivating PL_Alpha_Hill as a stable building block for our spectral signatures (Sec. 4.1). PL_Alpha_Hill also measures the dynamics of the ESD shape, and previous works (Yang et al., 2023; Liu et al., 2024) have shown that PL_Alpha_Hill reflects the training quality of different layers. Layers with larger PL_Alpha_Hill tend to be "over-trained," while layers with smaller PL_Alpha_Hill tend to be relatively "under-trained."

[1]It is known that more mathematically principled estimators exist, such as those based on MLE (Alstott et al., 2014; Clauset et al., 2009) or eigenvalue repulsion (Hodgkinson et al., 2025). However, in our experiments, the Hill estimator demonstrates more stable performance and is easier to scale to large settings.

Robustness of ESD Shape Metrics in Post-training. A critical prerequisite for a spectral signature to serve as a reliable lineage identifier is its robustness across downstream adaptations. We validate the robustness of the layer-wise PL_Alpha_Hill as a spectral signature by examining its stability under post-training transformations. We observe that the impact of post-training on the weight ESD is minimal; the spectral signature is predominantly shaped during the pre-training phase. As shown in Fig. 1, the spectral signature exhibits distinct behaviors for intra-family and inter-family comparisons. We observe a near-perfect spectral overlap between Llama-3.1 and its post-trained Tulu-3 variant. Simultaneously, the separation from the Llama-2 baseline is clear. This distinction supports that the spectral signature is primarily determined during pre-training and remains robust to post-training adjustments. We provide additional results in Appendix 7.5 to further corroborate these findings.

4. Method

Using robustness properties of PL_Alpha_Hill established in Sec. 3, we present our methodology for applying PL_Alpha_Hill-based spectral signatures to model quantification. As illustrated in Fig. 2, for a given collection of LLMs, we extract the PL_Alpha_Hill statistic from each weight matrix and aggregate these measurements to construct a model-level spectral signature, as formalized in Sec. 4.1. We then use these spectral signatures to address three model quantification tasks: (1) measuring similarity between LLMs, (2) model classification and clustering, and (3) predicting downstream LLM performance. The detailed methodology is shown in Sec. 4.2.

4.1. Formulation of Spectral Signatures

For a model f_θ in parameter space Θ with L transformer blocks and N_{mod} modules, we quantify the model by constructing a spectral signature: $\mathbf{Z} \in \mathbb{R}^{N_{\text{mod}} \times L}$, which is a tensor consisting of the PL_Alpha_Hill measure of each

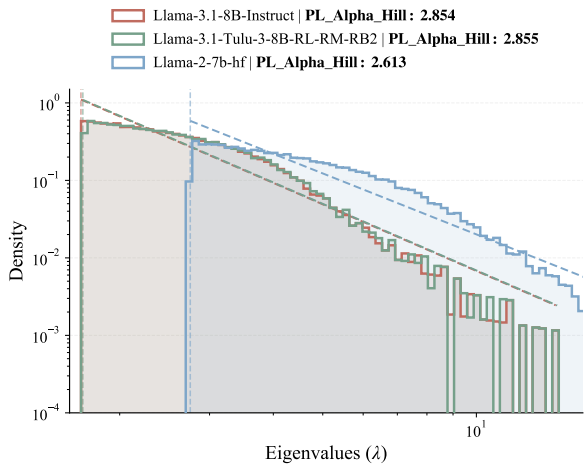

Figure 1. Eigenspectrum shape serves as a distinct signature for model lineage. We visualize the weight ESDs of three LLMs. Models sharing the same backbone (Llama-3.1 and its post-trained variant Tulu-3) exhibit nearly identical spectral shapes and power-law exponents ($PL_Alpha_Hill \approx 2.85$), while a model from a different family (Llama-2) shows significant deviation. This structural persistence validates the use of ESD shape as a robust spectral signature.

module. Each row $\mathbf{Z}_{i,:} \in \mathbb{R}^L$ is a vector of PL_Alpha_Hill of modules of the same type. For a standard transformer architecture, we have $N_{\text{mod}} = 7$. The tensor representation of spectral signature creates a mapping $\Theta \to \mathbb{R}^{N_{\text{mod}} \times L}$ that extracts spectral information of each weight matrix in the model. It allows module-level analysis, as different types of transformer layers often undergo different degrees of structural changes during post-training. Our empirical observation shows that Multi-Layer Perceptron (MLP) projection layers present larger separation in independently trained models compared to attention layers.

4.2. Spectral Signatures for Model Quantification

The spectral signature encodes robust spectral information of each weight matrix in an LLM, which can be used for systematic management and quantification of large collections of LLMs. In this section, we describe how spectral signatures can be systematically applied to model quantification tasks, and we demonstrate their use in the following representative scenarios.

Measuring LLM Similarities. To quantify the relations between models in a model collection $\mathcal{F}$, we define a similarity measure based on their spectral signatures. Let $\mathbf{Z}^{(k)} \in \mathbb{R}^{N_{\text{mod}} \times L}$ denote the spectral signature of model k, where the i-th row $\mathbf{Z}^{(k)}_{i,:} \in \mathbb{R}^L$ corresponds to the vector of PL_Alpha_Hill of all occurrences of the i-th module type (e.g., Attention Q/K/V, MLP). Given two models $f_{\theta_1}, f_{\theta_2}$ with signatures $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$, we compute the pairwise similarity by averaging the Spearman rank correlations across aligned modules:

$$\text{Sim}(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}) = \frac{1}{N_{\text{mod}}} \sum_{i=1}^{N_{\text{mod}}} \rho_s(\mathbf{Z}^{(1)}_{i,:}, \mathbf{Z}^{(2)}_{i,:}), \tag{2}$$

where $\rho_s(\cdot, \cdot)$ is the Spearman's rank correlation coefficient:

$$\rho_s(\mathbf{u}, \mathbf{v}) = \text{corr}(\text{rank}(\mathbf{u}),\ \text{rank}(\mathbf{v}))\,. \tag{3}$$

We use Spearman correlation in Eqs. 2–3 because it depends only on the relative depth-wise ordering of PL_Alpha_Hill values, making the similarity less sensitive to module-specific affine distortions of the statistics and improving comparability across scales. More importantly, PL_Alpha_Hill itself provides exact invariances that directly match our output-invariant stress tests. For any weight matrix $\mathbf{W}$, let $\mathbf{C} = \mathbf{W}^\top \mathbf{W}$ with eigenvalues $\{\lambda_j\}$. If $\mathbf{W}' = c\,\mathbf{P}\mathbf{W}\mathbf{Q}$ where $c > 0$ and $\mathbf{P}, \mathbf{Q}$ are orthogonal (including permutations), then $\mathbf{W}'^\top \mathbf{W}' = c^2\,\mathbf{Q}^\top \mathbf{C}\mathbf{Q}$ has eigenvalues $\lambda'_j = c^2 \lambda_j$. Since the Hill estimator used for PL_Alpha_Hill depends on log-ratios of tail eigenvalues, these ratios are unchanged and thus $PL_Alpha_Hill(\mathbf{W}') = PL_Alpha_Hill(\mathbf{W})$. This yields permutation- and scale-robust similarity without solving explicit neuron-matching problems that arise in matrix-level fingerprints such as AWM (Zeng et al., 2025).

Model Classification The properties of the spectral signature further enable its use for identifying the lineage of an LLM among candidate base models, as models derived from the same base model exhibit highly similar eigenspectrum shapes, whereas models with different origins show substantial divergence in shape. We approach the lineage identification as an unsupervised classification problem in the spectral signature space, and we use a nearest-neighbor classifier based on the spectral similarity from Eq. 2. Let $\{\mathbf{Z}^{(k)}\}_{k=1}^K$ denote the set of spectral signatures for K known base foundation models. For a given query model with spectral signature Z, we compute its similarity to each base model via the spectral signature similarity formula defined in Eq. 2. The predicted lineage $\hat{k}$ is the base model that maximizes this similarity:

$$\hat{k} = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \text{Sim}(\mathbf{Z}, \mathbf{Z}^{(k)}) \tag{4}$$

Model Clustering. Spectral signature of LLMs further enables the discovery of underlying genealogical structures within large model collections without supervision. We formulate this problem as a clustering problem in the space of spectral signatures. Given a model collection $\mathcal{F} = \{f_{\theta_1}, \ldots, f_{\theta_N}\}$, we extract the spectral signature $\mathbf{Z}^{(k)} \in \mathbb{R}^{N_{\text{mod}} \times L}$ of each model f_{θ_k}. For vector-based clustering algorithms, we flatten $\mathbf{Z}^{(k)}$ into a feature vector

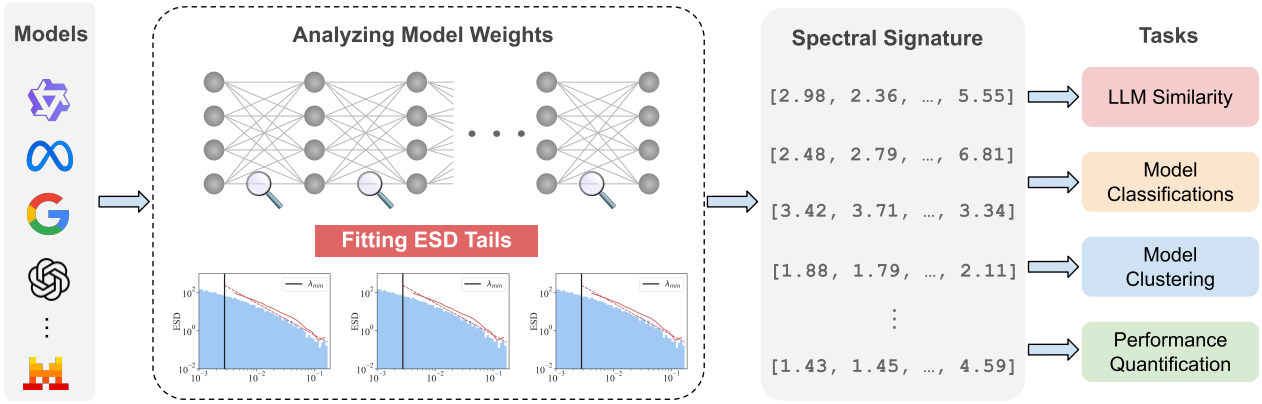

Figure 2. Overview of our proposed shape-based spectral signature framework. We analyze the eigenspectrum of model weights and use the shape information of ESDs (PL_Alpha_Hill) as spectral signatures for downstream tasks.

$\mathbf{z}^{(k)} \in \mathbb{R}^{N_{\text{mod}} \cdot L}$. We use spectral signatures for four unsupervised pipelines: HDBSCAN, Bayesian Gaussian Mixture Models (BGMM), Spectral Clustering, and K-Means. Due to the high dimensionality of the spectral signature representations, we apply Uniform Manifold Approximation and Projection (UMAP) as a preprocessing step for HDBSCAN, BGMM, and K-Means. In contrast, Spectral Clustering operates directly on the original high-dimensional pairwise distance matrix, where dimensionality reduction is implicitly handled through the graph Laplacian eigenstructure. Hyperparameters for clustering algorithms (i.e., number of clusters) are selected by maximizing the silhouette score. To ensure that clustering quality reflects intrinsic structure in the weight space rather than artifacts introduced by the embedding process, all internal evaluation metrics (e.g., silhouette scores) are computed using distances in the original spectral signature space. We assess the robustness and generalizability of each clustering pipeline by repeating experiments under random subsampling at varying sample sizes, verifying consistent clustering behavior across regimes.

Predicting LLM Performance. To quantify the predictive power of the proposed spectral signature, we treat the performance estimation as a non-parametric regression task over the model manifold. Given a query model f_θ with spectral signature $\mathbf{Z}$ and a reference pool of models $\mathcal{P} = \{(\mathbf{Z}^{(i)}, \mathbf{b}_i)\}_{i=1}^{N}$, where $\mathbf{Z}^{(i)}$ is the spectral signature and $\mathbf{b}_i$ is the vector of benchmark scores, we estimate the performance of f_θ using distance-weighted k-nearest neighbor (k-NN) Interpolation. Specifically, we first derive a distance metric $d(\mathbf{Z}, \mathbf{Z}^{(i)})$ from the spectral similarity $\text{Sim}(\mathbf{Z}, \mathbf{Z}^{(i)})$. We employ the transformation $d(\mathbf{Z}, \mathbf{Z}^{(i)}) = 1 - \text{Sim}(\mathbf{Z}, \mathbf{Z}^{(i)})$. This ensures that models with highly correlated spectral shapes are positioned closer in the embedding space. The predicted benchmark scores $\tilde{\mathbf{b}}$ are computed by aggregating the top-k most similar neighbors in $\mathcal{P}$:

$$\tilde{\mathbf{b}} = \frac{\sum_{i \in \mathcal{K}} w_i \cdot \mathbf{b}_i}{\sum_{i \in \mathcal{K}} w_i},$$

, where $\mathcal{K}$ is the set of k nearest neighbors, and $w_i = 1/(d(\mathbf{Z}, \mathbf{Z}^{(i)}) + \epsilon)$ is the assigned weight. This interpolation relies on the assumption that models that exhibit similar HT-SR patterns in their weight matrices are likely to possess comparable generalization capabilities. We then quantify the prediction accuracy with:

$$D_{perf} = \|\tilde{\mathbf{b}} - \mathbf{b}\|, \tag{5}$$

where $\mathbf{b}$ represents the ground truth benchmark scores. Typically, $\|\cdot\|$ in Eq. 5 is defined by mean absolute error (MAE).

5. Experiments

In this section, we demonstrate that spectral signatures provide an efficient and accurate tool for a range of downstream tasks. Specifically, we present details of the experimental setup as well as the results of (i) measuring LLM similarities, (ii) model classification and clustering, and (iii) generalization to diverse architectures.

5.1. Measuring LLM Similarities

Model Corpus. We select Llama-2-7b (Touvron et al., 2023b) as the base model and compare it with four types of related variants that share the same architecture: (i) post-trained models (SFT and RL), (ii) merged models and pruned models, (iii) noise-perturbed models obtained by adding i.i.d. Gaussian perturbations to all weight entries with noise strength $\gamma \in \{0.05, 0.30, 1.00\}$, as defined in Appendix 7.9, (iv) output-invariant transformations on adjacent matrices and output-invariant permutations of hidden units, as defined in Appendix 7.9. This corpus isolates a key practical goal: a lineage/similarity signal should remain

stable under post-training and output-invariant reparameterizations, while being informative under destructive edits (e.g., aggressive pruning). We further validate the generalizability of our spectral signature to Mixture-of-Experts architectures and cross-depth model pairs in Appendix 7.8

Setup. We compare against both *data-aware* and *data-free* baselines. For data-aware methods, we use REEF (Zhang et al., 2025), which computes similarity from activation features on a fixed prompt set (300 prompts for TruthfulQA (Lin et al., 2022)), and LOGITS (Yang & Wu, 2024), which measures similarity from model outputs on the same prompts. For data-free approaches, we apply PCS (Zeng et al., 2024) that computes cosine similarity between flattened weight tensors under a fixed alignment, and discuss alignment-based weight signature (e.g., AWM (Zeng et al., 2025)) as a complementary point of comparison: these methods operate at the matrix level and often require per-pair matching/alignment, which makes them less convenient as a drop-in kernel when building dense similarity matrices for zoo-scale clustering and retrieval.

Results. Tab. 2 summarizes the overall similarity performance of our method compared to baseline models. It demonstrates the robustness of our spectral signature under diverse scenarios, including fine-tuning, pruning, model merging, noise injection, and output-invariant attacks. First, the similarity scores of our method remain consistently high among Llama-2-7b related variants, such as multiple fine-tuned models, with scores in the $[0.97, 1.0]$ range across instruction tuning and domain shifts. In contrast, data-aware baselines can fail under prompt- or language-induced behavior changes. For example, Logits drops sharply for the Chinese variants, and REEF degrades for the LoRA-unloaded model, whereas our scores remain stable. Under weight noise injection, our similarity decreases mildly as σ increases from ≈ 0.999 at $\gamma = 0.05$ to ≈ 0.979 at $\gamma = 1.00$, outperforming PCS and REEF at the strongest noise perturbation. Moreover, our method is robust to hidden-unit permutations and random scaling under output-invariant transformations, while PCS is sensitive to both. For unstructured pruning, the similarity score using spectral signature decreases gradually at 30–50% sparsity and drops substantially at 70% sparsity, indicating that the spectral signature is informative of the structural changes under more aggressive parameter removal. Quantifying similarity among large model corpora typically requires constructing a stable and efficient pairwise similarity matrix. However, existing methods often fall short computationally: data-dependent methods require prompts and incur inference costs for each comparison, while pair-specific methods such as AWM incur cumulative overhead when constructing dense similarity matrices over large model collections. Our approach computes a compact signature once per model and yields a computationally-cheap similarity that remains robust under the varying transformations that break coordinate-dependent weight distances, making it practical for large-scale lineage analysis.

5.2. Model Classification and Clustering

Model Corpus. We construct a corpus of 128 publicly available models derived from 8 model families: llama-7b (Touvron et al., 2023a), llama-2-7b (Touvron et al., 2023b), llama-3.1-8b (Grattafiori et al., 2024), mistral-7b-v0.1, mistral-7b-v0.2 (Jiang et al., 2023), open-llama-7b, open-llama-7b-v2 (Geng & Liu, 2023), and qwen-1.5-7b (Bai et al., 2023). For each derived model, we assign a training-lineage label corresponding to one of the eight base models, using the model's publicly documented training provenance. For the classification task, we classify each of the 128 derived models into one of the 8 base families by selecting the base with the highest spectral signature similarity score to the model. For the clustering task, we perform unsupervised clustering on the signatures of all 128 derived models and 8 base models.

Setup. For the classification task, we compute the similarity score using three different kernels on spectral signatures: Spearman rank correlation, radial basis function (RBF) kernel, and cosine similarity (detailed in Appendix Sec. 7.2). This allows us to assess the robustness of classification performance under different similarity measures. For the clustering task, we conduct two complementary studies. First, we assess clustering performance using spectral signatures constructed with different ESD-based metrics, including shape metrics (PL_Alpha_Hill), scale metrics (Log_Norm and Spectral_Norm), and a combined shape–scale metric (Log_Alpha_Norm). Second, we evaluate the robustness of clustering by applying four clustering algorithms, as described in Sec. 4, to these spectral signatures. We assess clustering quality using the Silhouette Coefficient (Rousseeuw, 1987) and the Davies–Bouldin Index (Davies & Bouldin, 2009), as detailed in Appendix 7.3.

Classification Results. Fig. 7 shows that the proposed spectral signature enables accurate lineage classification across all eight model families with an overall accuracy of 98.44%. Across similarity kernels, the classifier consistently achieves near-perfect accuracy, indicating that lineage information is strongly preserved in the spectral signature. Performance is stable across Spearman rank, RBF, and cosine similarity, suggesting that the results are not sensitive to the choice of similarity measure. Minor accuracy variations are mainly observed for closely related families, such as mistral-7b-v0.1 and mistral-7b-v0.2, which correspond to different versions within the same model line and share highly similar architectures and training procedures. Overall, these results

Table 2. A comprehensive evaluation of signature-based methods for measuring LLM similarity across diverse settings. Here, , , and denote similarity scores > 0.8 (high), between 0.5 and 0.8 (moderate), and < 0.5 (low), respectively.

	Model Fine-tuning					
	llama2-7b-hf-instruction-lora	vicuna-v1.5	Nous-Hermes-llama-2-7b	chinese-llama-2-7b	Llama-2-7b-hf-instruct-pl-lora_unload	chinese-alpaca-2-7b
PCS (Zeng et al., 2024)	0.99	0.99	0.99	0.93	0.99	0.95
Logits (Yang & Wu, 2024)	0.98	0.79	0.99	0.03	0.99	0.56
REEF (Zhang et al., 2025)	0.77	0.99	0.99	0.99	0.14	0.99
GhostSpec (Wang et al., 2025)	0.98	0.99	0.99	0.80	0.87	0.83
Ours	0.99	0.99	0.99	0.97	0.99	0.97
	Unrelated Models					
	llama-7b	llama-3.1-8B	Mistral-7B	Qwen-1.5-7B	open-llama-7b	open-llama-7b-v2
PCS (Zeng et al., 2024)	0.01	0.14	0.27	0.06	0.01	0.01
Logits (Yang & Wu, 2024)	0.99	0.00	0.77	0.00	0.96	0.95
REEF (Zhang et al., 2025)	0.15	0.15	0.13	0.14	0.17	0.18
GhostSpec (Wang et al., 2025)	0.94	0.19	0.00	0.89	0.91	0.80
Ours	0.92	0.66	0.77	0.68	0.72	0.60
	Unstructured Pruning			**RL Models**		
	30%	50%	70%	Archer-Code-1.5B (GRPO)	Nemotron-Research-Reasoning-Qwen-1.5B (Reinforcement++)	Polaris-7B-Preview (DAPO)
PCS (Zeng et al., 2024)	0.99	0.90	0.78	0.99	0.99	1.00
Logits (Yang & Wu, 2024)	1.00	1.00	0.99	1.00	1.00	1.00
REEF (Zhang et al., 2025)	0.99	0.98	0.98	0.98	0.96	0.94
GhostSpec (Wang et al., 2025)	0.99	0.89	0.70	1.00	1.00	0.46
Ours	0.89	0.89	0.40	0.99	0.99	0.99
	Noise Injection			**Weight Merging (Evollm-jp)**		
	$\gamma = 0.05$	$\gamma = 0.30$	$\gamma = 1.00$	shisa-gamma-7b	wizardmath-7b-1.1	abel-7b-002
PCS (Zeng et al., 2024)	0.99	0.95	0.70	0.99	0.99	0.99
Logits (Yang & Wu, 2024)	1.00	1.00	0.97	0.00	1.00	1.00
REEF (Zhang et al., 2025)	0.99	0.99	0.69	0.93	0.91	0.89
GhostSpec (Wang et al., 2025)	1.00	0.97	0.70	0.23	0.99	0.99
Ours	0.99	0.98	0.97	0.80	0.87	0.83

demonstrate that spectral signatures provide a reliable and discriminative basis for model lineage classification.

Clustering Results. Tab. 3 compares the clustering performance of spectral signatures constructed from scale- and shape-based metrics. The proposed shape-based PL_Alpha_Hill achieves the best performance, with the highest Silhouette score of 0.91 and the lowest Davies-Bouldin index of 0.22, consistently outperforming the scale-based alternatives. To further evaluate the robustness of our proposed spectral signature, in Tab. 4 we report the result of different clustering algorithms. All clustering algorithms that use the spectral signature achieve high silhouette scores, ranging from 0.73 to 0.91. This shows that the proposed spectral signature provides a robust and effective embedding for different classes of clustering algorithms. As an illustrative example, in Fig. 8 we provide the clustering result of 2D t-SNE using the spectral clustering algorithm with spectral signatures constructed by PL_Alpha_Hill. As we can see, the proposed spectral signature clearly separates different model families, resulting in distinct clusters.

Table 3. Comparison of scale- and shape-based spectral signatures for spectral clustering.

Metric	Silhouette (↑)	Davies_Bouldin (↓)
Log_Norm (scale)	0.70	0.47
Spectral_Norm (scale)	0.81	0.35
Log_Alpha_Norm (scale)	0.88	0.23
PL_Alpha_Hill (shape)	**0.91**	**0.22**

Table 4. Comparing different clustering algorithms using the spectral signature constructed by PL_Alpha_Hill.

Method	Silhouette (↑)	Davies_Bouldin (↓)
HDBSCAN	0.73	0.56
BGMM	0.73	0.56
K-Means	0.73	0.56
Spectral Clustering	0.91	0.22

5.3. Quantifying and Ranking LLM Performance

Model Corpus. For model ranking and prediction tasks, we curate 499 open-source models from the Open LLM Leaderboard (Myrzakhan et al., 2024). This corpus spans a broad spectrum of architectures, focusing on the Llama, Mistral, GPT, and Pythia families, with parameter scales ranging from 19 million to 70 billion. Further validation on additional scales and intra-family prediction is provided in Appendix 7.8.

Setup. To evaluate whether spectral signatures can predict downstream performance, we consider four representative benchmarks: ARC-Challenge (reasoning) (Clark et al., 2018), HellaSwag (commonsense) (Zellers et al., 2019), MMLU (world knowledge) (Hendrycks et al., 2021), and TruthfulQA (factuality) (Lin et al., 2022). The models span

a wide performance range (e.g., MMLU scores from 23.0 to 70.67), covering both low- and high-performance regimes. For each model, we estimate performance using distance-weighted k-NN interpolation ($k = 3$) based on spectral-signature similarity. We use 5-fold cross-validation to assess generalization.

We compare against two activation-based baselines. First of all, EmbedLLM (Zhuang et al., 2024) encodes each model by extracting the hidden-state embedding at the final token position from the last transformer layer on a fixed prompt set $\mathcal{P}$, and trains a linear regressor to predict benchmark scores under leave-one-out evaluation. Second, LLMDNA (Wu et al., 2025) extracts a 128×128 model representation from 100 random samples, which is nearly invariant to the specific dataset or questions used for extraction. We use the same regression protocol for LLMDNA as for EmbedLLM. Prediction quality is evaluated by MAE and Kendall's Tau rank correlation (Kendall, 1948). We define "Significance" as the number of repeated train/test splits (out of 100), where Kendall's Tau between predicted and ground-truth rankings is statistically significant at the 5% level. In Appendix 7.7, we report that our method is significantly computationally cheaper than the baselines.

We further discuss why our spectral signatures are well suited for this setting. Activation-based performance quantification requires task data at inference, and does not scale well to frequent re-ranking across large model collections. Conversely, other weight-based methods such as PCS (Zeng et al., 2024) and AWM (Zeng et al., 2025) are primarily designed for pair-wise provenance checks and do not naturally provide a compact, geometry-aware representation amenable to regression across heterogeneous architectures. Our spectral signatures provide a lightweight, task-agnostic representation that supports nearest-neighbor interpolation without any prompts.

Results. We run 100 repeated trials with random train/test splits (249 models for training and 250 models for testing in each trial). Fig. 9 compares predicted benchmark scores with their ground-truth values. The points cluster closely around the diagonal ($y = x$), indicating good agreement between predicted and actual performance. We report (i) the average MAE over the 100 trials, which measures absolute prediction error, and (ii) the number of trials in which Kendall's Tau correlation between predicted and ground-truth rankings on the test set is statistically significant. As shown in Tab. 5, spectral signatures achieve consistently low MAE and statistically significant rank correlation in all trials across all benchmarks. While EmbedLLM achieves slightly lower MAE on ARC and TruthfulQA, spectral signatures outperform EmbedLLM on MMLU and remain competitive on HellaSwag. This gap reflects that benchmarks targeting specific model capabilities (TruthfulQA, ARC) are more directly reflected in task-specific activations, while benchmarks covering general knowledge (MMLU, HellaSwag) align more closely with the pretraining capacity encoded in weight geometry. LLMDNA (Wu et al., 2025), despite requiring comparable inference cost, yields substantially higher MAE across all benchmarks. Unlike EmbedLLM, which extracts task-conditioned activations, LLMDNA produces a single task-agnostic embedding per model. The compression loses the benchmark-specific variability needed for accurate performance prediction.

Table 5. Performance prediction results on four benchmarks, comparing our spectral signatures, EmbedLLM, and LLMDNA using MAE and statistical significance (Signif.).

	Ours		**EmbedLLM**		**LLMDNA**	
Benchmark	**MAE**	**Signif.**	**MAE**	**Signif.**	**MAE**	**Signif.**
ARC	2.27	100	1.61	100	7.12	100
HellaSwag	3.25	100	3.50	100	11.65	100
MMLU	1.95	100	3.14	100	6.78	100
TruthfulQA	2.99	100	1.96	100	3.43	100

6. Conclusion

This work presents a spectral shape–based framework for large-scale analysis of LLMs that is data-free, computationally efficient, and robust to post-training transformations. We curate and benchmark a large and diverse LLM corpus and conduct extensive experiments across similarity measurement, classification, clustering, and performance prediction. Our results have shown the consistent effectiveness of our spectral signature across various experimental setups. Furthermore, we validate that ESD shape metrics, particularly PL_Alpha_Hill, capture intrinsic structural properties of LLMs that are stable under output-invariant reparameterizations and noise perturbations. Therefore, our spectral signature can provide a practical tool for managing modern LLM ecosystems, with implications for model provenance, organization, and scalable evaluation.

Acknowledgements

YY would like to acknowledge the support of the U.S. Department of Energy (Grant number: DE-SC0025584) and the Defense Advanced Research Projects Agency (Grant number: HR00112520011). HL was supported by the U.S. Department of Energy under Contract DE-AC02-05CH11231 and U.S. National Science Foundation NSF-DMS 2412403. We also would like to acknowledge NERSC DOE Mission Science Allocation for ERCAP request ERCAP0035003.

References

Alstott, J., Bullmore, E., and Plenz, D. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014.

Bai, J., Bai, S., Chu, Y., et al. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. *SIAM review*, 51(4): 661–703, 2009.

Couillet, R. and Liao, Z. *Random matrix methods for machine learning*. Cambridge University Press, Cambridge, UK, 2022.

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-1(2):224–227, 2009.

Geng, X. and Liu, H. Openllama: An open reproduction of llama, May 2023. URL https://github.com/openlm-research/open_llama.

Grattafiori, A., Dubey, A., Jauhri, A., et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gu, C., Huang, C., Zheng, X., Chang, K.-W., and Hsieh, C.-J. Watermarking pre-trained language models with backdooring, 2022. URL https://api.semanticscholar.org/CorpusID:252907247.

Gueta, A., Venezian, E., Raffel, C., Slonim, N., Katz, Y., and Choshen, L. Knowledge is a region in weight space for finetuned language models, 2023.

Gurbuzbalaban, M., Simsekli, U., and Zhu, L. The heavy-tail phenomenon in sgd. In *International Conference on Machine Learning*, pp. 3964–3975, Virtual, 2021. PMLR.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Hill, B. M. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5): 1163–1174, 1975.

Hodgkinson, L. and Mahoney, M. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274, Virtual, 2021. PMLR.

Hodgkinson, L., Simsekli, U., Khanna, R., and Mahoney, M. Generalization bounds using lower tail exponents in stochastic optimizers. In *International Conference on Machine Learning*, pp. 8774–8795, Baltimore, Maryland, USA, 2022. PMLR.

Hodgkinson, L., Wang, Z., and Mahoney, M. W. Models of heavy-tailed mechanistic universality, 2025.

Horwitz, E., Kurer, N., Kahana, J., Amar, L., and Hoshen, Y. We should chart an atlas of all the world's models, 2025a.

Horwitz, E., Shul, A., and Hoshen, Y. Unsupervised model tree heritage recovery. In *Proceedings of the 13th International Conference on Learning Representations*, pp. 47900–47918, Singapore, 2025b. OpenReview.net.

Hu, Y., Goel, K., Killiakov, V., and Yang, Y. Eigenspectrum analysis of neural networks without aspect ratio bias. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 24290–24313, Vancouver, Canada, 2025. PMLR.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jiang, Y., Wu, C., Boroujeny, M. K., Mark, B., and Zeng, K. Stealthink: A multi-bit and stealthy watermark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 27685–27709, Vancouver, Canada, 2025. PMLR.

Jin, H., Zhang, C., Shi, S., Lou, W., and Hou, Y. T. Proflingo: A fingerprinting-based intellectual property protection scheme for large language models. In *2024 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–9, Taipei, Taiwan, 2024. IEEE.

Kendall, M. G. *Rank Correlation Methods*. Charles Griffin & Company, London, UK, 1948.

Liao, Z. and Mahoney, M. W. Random matrix theory for deep learning: Beyond eigenvalues of linear models, 2025.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics.

Liu, Z., Hu, Y., Pang, T., Zhou, Y., Ren, P., and Yang, Y. Model balancing helps low-data training and fine-tuning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1311–1331, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.78. URL https://aclanthology.org/2024.emnlp-main.78/.

Lu, H., Zhou, Y., Liu, S., Wang, Z., Mahoney, M. W., and Yang, Y. Alphapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. *Advances in neural information processing systems*, 37:9117–9152, 2024.

Mahoney, M. and Martin, C. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pp. 4284–4293, Long Beach, CA, USA, 2019. PMLR.

Martin, C. H. and Hinrichs, C. Setol: A semi-empirical theory of (deep) learning, 2025.

Martin, C. H. and Mahoney, M. W. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Martin, C. H., Peng, T. S., and Mahoney, M. W. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):1–13, 2021.

Myrzakhan, A., Mahmoud Bsharat, S., and Shen, Z. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024.

Nikolic, I., Baluta, T., and Saxena, P. Model provenance testing for large language models. In *Advances in Neural Information Processing Systems*, pp. 34126–34153, Sydney, Australia, 2025. Curran Associates, Inc.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Schürholt, K., Meynent, L., Zhou, Y., Lu, H., Yang, Y., and Borth, D. A model zoo on phase transitions in neural networks, 2025.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837, Long Beach, CA, USA, 2019. PMLR.

Simsekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

Sun, M., Yin, Y., Xu, Z., Kolter, J. Z., and Liu, Z. Idiosyncrasies in large language models, 2025.

Tang, Y., Li, X., Liu, F., Guo, W., Zhou, H., Wang, Y., Han, K., Yu, X., Li, J., Zang, H., Mi, F., Meng, X., Liu, Z., Chen, H., Zheng, B., Chen, C., Yan, Y., Tang, R., Qin, P., Chen, X., Tao, D., and Wang, Y. Pangu pro moe: Mixture of grouped experts for efficient sparsity, 2025. URL https://arxiv.org/abs/2505.21411.

Touvron, H., Lavril, T., Izacard, G., et al. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Touvron, H., Martin, L., Stone, K., et al. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Uchida, Y., Nagai, Y., Sakazawa, S., and Satoh, S. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ICMR '17, pp. 269–277, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450347013. doi: 10.1145/3078971.3078974. URL https://doi.org/10.1145/3078971.3078974.

Wang, S., Ma, Z., Li, X., and Li, Z. Ghost in the transformer: Tracing llm lineage with svd-fingerprint, 2025.

Wang, T. and Kerschbaum, F. Attacks on digital watermarks for deep neural networks. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2622–2626, Brighton, UK, 2019. IEEE. doi: 10.1109/ICASSP.2019.8682202.

Wu, K., Pang, L., Shen, H., Cheng, X., and Chua, T.-S. LLMDet: A third party large language models generated text detection tool. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2113–2133, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.139. URL https://aclanthology.org/2023.findings-emnlp.139/.

Wu, Z., Zhao, H., Wang, Z., Guo, J., Wang, Q., and He, B. Llm dna: Tracing model evolution via functional representations, 2025.

Yang, Y., Hodgkinson, L., Theisen, R., Zou, J., Gonzalez, J. E., Ramchandran, K., and Mahoney, M. W. Taxonomizing local versus global structure in neural network loss

landscapes. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18722–18733, Virtual, 2021. Curran Associates, Inc.

Yang, Y., Theisen, R., Hodgkinson, L., Gonzalez, J. E., Ramchandran, K., Martin, C. H., and Mahoney, M. W. Test accuracy vs. generalization gap: Model selection in nlp without accessing training or testing data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3011–3021, Long Beach, CA, USA, 2023. ACM.

Yang, Z. and Wu, H. A fingerprint for large language models, 2024. URL https://api.semanticscholar.org/CorpusID:270869924.

Yoon, D.-h., Chun, M., Allen, T., Müller, H., Wang, M., and Sharma, R. Intrinsic fingerprint of llms: Continue training is not all you need to steal a model!, 2025.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Zeng, B., Wang, L., Hu, Y., Xu, Y., Zhou, C., Wang, X., Yu, Y., and Lin, Z. HuRef: HUman-REadable fingerprint for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126332–126362, Vancouver, Canada, 2024. Curran Associates, Inc.

Zeng, B., Chen, L., He, Z., Wang, X., and Lin, Z. Awm: Accurate weight-matrix fingerprint for large language models, 2025.

Zhang, J., Liu, D., Qian, C., Zhang, L., Liu, Y., Qiao, Y., and Shao, J. REEF: representation encoding fingerprints for large language models. In *International Conference on Learning Representations*, pp. 48092–48117, Singapore, 2025. OpenReview.net.

Zhou, Y., Pang, T., Liu, K., Mahoney, M. W., Yang, Y., et al. Temperature balancing, layer-wise weight analysis, and neural network training. *Advances in Neural Information Processing Systems*, 36:63542–63572, 2023.

Zhu, S., Ahmed, A. M., Kuditipudi, R., and Liang, P. Independence tests for language models. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 79673–79698, Vancouver, Canada, 2025. PMLR.

Zhuang, R., Wu, T., Wen, Z., Li, A., Jiao, J., and Ramchandran, K. Embedllm: Learning compact representations of large language models, 2024. URL https://arxiv.org/abs/2410.02223.

7. Appendix

7.1. Metric Computation

In this section, we provide the detailed definitions of the metrics used in our analysis. Let $\mathbf{W}$ represent the weight matrix.

Log_Norm. Log_Norm is defined as the logarithm of the squared Frobenius norms:

$$\text{Log_Norm} = \log \|\mathbf{W}\|_F^2. \tag{6}$$

Log_Alpha_Norm. This metric is a scale-adjusted metric based on the Schatten p-norm. Let $\{\lambda_j\}_{j=1}^M$ denote the set of eigenvalues of the correlation matrix $\mathbf{X} = \mathbf{W}^\top \mathbf{W}$, where $\mathbf{W}$ is an $N \times M$ weight matrix (assuming $N \geq M$). The Schatten p-norm is defined as:

$$\|\mathbf{X}\|_p = \left(\sum_{j=1}^{M} \lambda_j^p \right)^{\frac{1}{p}}. \tag{7}$$

The Log_Alpha_Norm metric is then given by:

$$\text{Log_Alpha_Norm} = \log \|\mathbf{X}\|_\alpha^\alpha. \tag{8}$$

Spectral_Norm. The Spectral_Norm focuses on the spectral properties of the weight matrices. Unlike the logarithmic version, here we compute the average squared spectral norm (operator norm) across all layers:

$$\text{Spectral_Norm} = \|\mathbf{W}\|_2^2, \tag{9}$$

where $\|\mathbf{W}\|_2$ denotes the spectral norm of the matrix $\mathbf{W}$, which corresponds to its largest singular value.

7.2. Metrics for Classification Quality

Given two signatures $\mathbf{Z}_1, \mathbf{Z}_2$, we consider: (i) Spearman rank correlation $\rho_s(\mathbf{Z}_1, \mathbf{Z}_2)$ as our default similarity (Eq. 3), (ii) an RBF kernel applied to the signature distance (Eq. 10),

$$k_{\text{RBF}}(\mathbf{Z}_1, \mathbf{Z}_2) = \exp\left(-\frac{\|\mathbf{Z}_1 - \mathbf{Z}_2\|_2^2}{2\sigma^2} \right) \tag{10}$$

, and (iii) cosine similarity (Eq. 11).

$$k_{\cos}(Z_1, \mathbf{Z}_2) = \frac{\langle \mathbf{Z}_1, \mathbf{Z}_2 \rangle}{\|\mathbf{Z}_1\|_2 \|\mathbf{Z}_2\|_2} \tag{11}$$

7.3. Metrics for Internal Clustering Validation

In the absence of ground truth labels for the unsupervised discovery of model families, we rely on internal validity indices to evaluate the quality of the clustering structures derived from spectral signatures. We employ two standard metrics: the Silhouette Coefficient and the Davies-Bouldin Index. Detailed definitions are provided below.

Silhouette Coefficient The Silhouette Coefficient evaluates the quality of clustering by measuring how similar an object is to its own cluster (cohesion) compared to other clusters (separation). For a data point $i \in C_k$ (where C_k is a given cluster), let $a(i)$ be the average distance between i and all other points in the same cluster, and let $b(i)$ be the minimum average distance from i to all points in any other cluster C_j (where $j \neq k$). The silhouette coefficient $s(i)$ is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{12}$$

The overall Silhouette Score is the mean of $s(i)$ across all samples.

Davies-Bouldin Index The Davies-Bouldin Index (DBI) signifies the average 'similarity' between clusters, where similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

Let S_i be the intra-cluster dispersion (e.g., average distance of points to the centroid) of cluster i, and let M_{ij} be the distance between the centroids of clusters i and j. The similarity ratio R_{ij} is defined as:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{13}$$

The Davies-Bouldin Index is defined as the average of the maximum similarity ratios for each cluster:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} R_{ij} \tag{14}$$

7.4. Similarity within Modules

Fig. 3 shows module-wise Spearman similarity heatmaps computed from layer-wise PL_Alpha_Hill profiles. A clear contrast emerges between attention and MLP modules. For the attention Q projection, many cross-family pairs remain highly similar (e.g., open-llama-7b vs. llama-2-7b reaches $\text{Sim} \approx 0.90$), resulting in a relatively weak separation between related and unrelated models. In contrast, the MLP up projection module exhibits a much sharper block-wise structure, where models derived from the same base family form high-similarity blocks, while similarities across families drop substantially.

We hypothesize that this difference reflects the roles of the two components in Transformer weights. Attention projections are strongly constrained by shared architectural parameterization and generic token-mixing behavior, which can lead to similar spectral shapes even across independently

trained models of comparable scale. MLP projections, however, are more directly responsible for feature transformation and information storage, and thus may encode more lineage-specific training signals, producing clearer family-level clusters in spectral space.

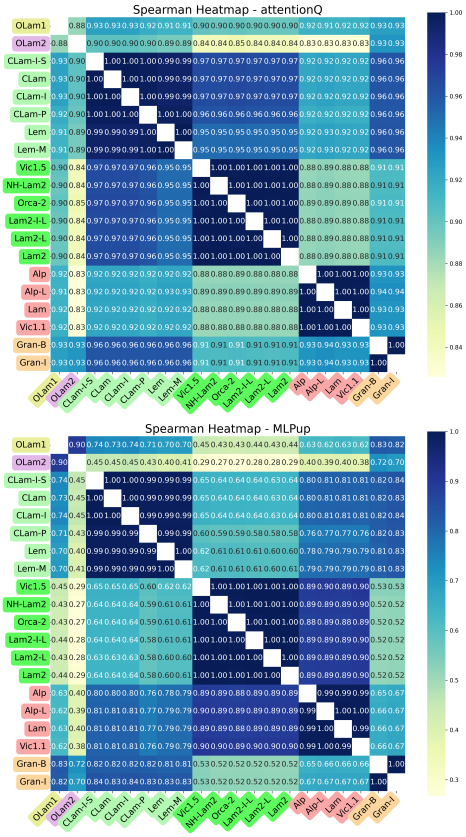

Figure 3. Module-wise spectral similarity heatmaps. PL_Alpha_Hill-SR signatures computed separately for (top) the attention Q projection and (bottom) the MLP projection module.

7.5. Additional Details on the Robustness of ESD Shape Metrics

Pre-training and post-training processes inherently reshape the correlations within weight matrices, which are reflected in the variations of the heavy-tail index across layers. To validate the reliability of our method, we investigate whether the spectral structure preserves sufficient lineage information despite these post-training perturbations.

PL_Alpha_Hill is robust under perturbation. Fig. 4 shows that ESD shape metrics can encode lineage information beyond pretraining. Using Spearman correlation between layer-wise PL_Alpha_Hill profiles, models derived from the same 7B base checkpoint form clear high-similarity blocks, whereas similarities across base families are substantially lower. Moreover, within the Llama-2 lineage, we observe finer sub-structure: distinct clusters align with different post-training trajectories (e.g., instruction-tuned vs. CodeLlama-style derivatives), suggesting that post-training alters spectral signatures in a structured manner. These observations motivate our use of spectral signatures for downstream tasks such as lineage attribution, clustering, and performance prediction (Sec.5).

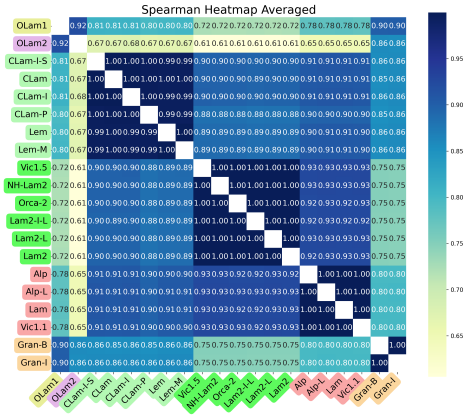

Figure 4. Heatmap of spectral signature similarity across models derived from different 7B base models. The pairwise similarity is measured by Spearman's rank correlation of PL_Alpha_Hill. Entries are color-coded on the axes to indicate the base model. Specifically, different shades of green represent models derived from Llama-2-7b. While dark green denotes the ones directly fine-tuned from the Llama-2-7B family, light green identifies the CodeLlama-7B-hf sub-family. The emergence of distinct clusters aligned with these color markers demonstrates that spectral signatures are highly sensitive to ancestral relations, even after diverse fine-tuning trajectories. We provide the full name list for the abbreviations in Tab. 6.

The Forging of Spectral Signature As illustrated in Fig. 5, we observe three distinct phases of spectral evolution:

1. **Initialization.** At Step 0, all layers exhibit large, uniform PL_Alpha_Hill values ($\alpha \approx 2.1$ for attention weight matrices and $\alpha \approx 3.5$ for MLP weight matrices, due to the aspect ratio of weight matrices (Hu et al., 2025)), resembling random matrices with no structural identity.

2. **Formulation.** As training progresses, α decreases globally, signifying the accumulation of information. Crucially, a distinct layer-wise topology emerges: PL_Alpha_Hill drop significantly in early layers (feature extraction), flatten in middle layers, dip in mid-deep layers, and rise in the final layers.

3. **Stabilization** As shown in the temporal similarity heatmap (Fig. 6), the spectral signature stabilizes rapidly. While the early phase ($< 16k$ steps) shows high volatility, the signature "locks in" after step 16,000. Subsequent updates (from 16k to 143k) refine

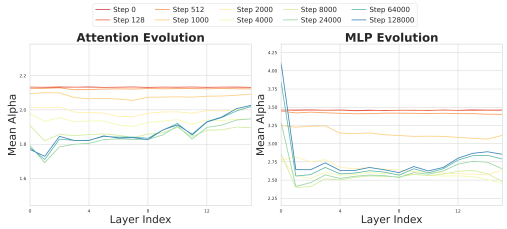

Figure 5. The evolution of layer-wise PL_Alpha_Hill through the training process. The values of PL_Alpha_Hill are averaged within Attention modules and MLP modules.

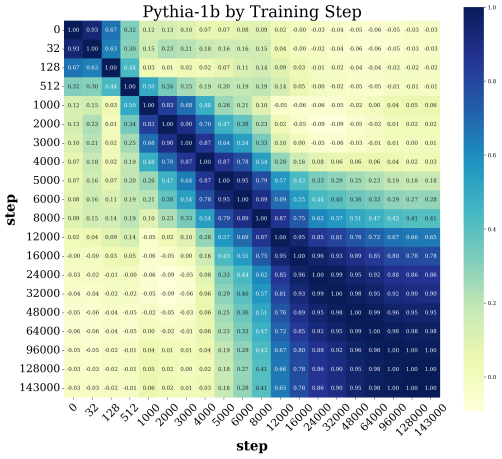

Figure 6. The heatmap of the Spearman rank between PL_Alpha_Hill signatures of pythia-1b models with training steps ranging from step 0 to step 143,000.

the weights but do not alter the fundamental topological shape.

PL_Alpha_Hill undergo minimal change during post-training. Because post-training updates (e.g., supervised fine-tuning and reinforcement learning from human feedback) typically introduce relatively small changes in parameter space, they are unlikely to erase the heavy-tailed structure established during pre-training. Based on that, PL_Alpha_Hill is promising to work as a stable and lineage-preserving signature: it is sensitive to the underlying architecture and pre-training data, but remains robust to common downstream adaptations and moderate noise. This motivates our use of ESD-based metrics for both performance prediction and model attribution.

7.6. Additional Results on Model Clustering and Classification

In Fig. 7, we demonstrate the accuracy of the 8-way classifier across families.

In Fig. 8, we show a t-SNE visualization of a 2D embedding for spectral clustering on PL_Alpha_Hill with ground truth cluster labels.

Table 6. Mapping of Model Abbreviations to Full Model Names

Abbreviation	Full Model Name
OLam1	`openlm-research/open_llama_7b`
OLam2	`openlm-research/open_llama_7b_v2`
CLam-I-S	`AlfredPros/CodeLlama-7b-Instruct-Solidity`
CLam	`codellama/CodeLlama-7b-hf`
CLam-I	`codellama/CodeLlama-7b-Instruct-hf`
CLam-P	`codellama/CodeLlama-7b-Python-hf`
Lem	`EleutherAI/llemma_7b`
Lem-M	`EleutherAI/llemma_7b_muinstruct_camelmath`
Vic1.5	`lmsys/vicuna-7b-v1.5`
NH-Lam2	`NousResearch/Nous-Hermes-llama-2-7b`
Orca-2	`microsoft/Orca-2-7b`
Lam2-I-L	`lvkaokao/llama2-7b-hf-instruction-lora`
Lam2-L	`oh-yeontaek/llama-2-7B-LoRA-assemble`
Lam2	`meta-llama/Llama-2-7b-hf`
Alp	`wxjiao/alpaca-7b`
Alp-L	`chainyo/alpaca-lora-7b`
Lam	`luodian/llama-7b-hf`
Vic1.1	`lmsys/vicuna-7b-v1.1`
Gran-B	`ibm-granite/granite-7b-base`
Gran-I	`ibm-granite/granite-7b-instruct`

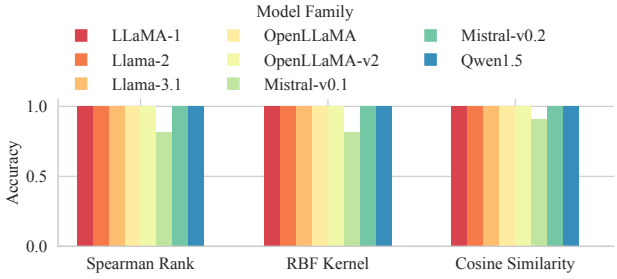

Figure 7. Classification accuracy of the 8-way classifier for routing derived models to their base models, reported across model families and similarity kernels. Each bar shows accuracy for one family-kernel pair.

7.7. Additional Results on Performance Predictions

Fig. 9 compares predicted benchmark scores with their ground-truth values. The points cluster closely around the diagonal ($y = x$), indicating good agreement between predicted and actual performance.

Computational Cost. We report a runtime comparison between our spectral signature and baselines under the same model and hardware settings. We use Llama-2-7b as the representative backbone and run all methods on a single NVIDIA L40 GPU with FP16 precision. For EmbedLLM, we follow the original pipeline and extract last-token hidden-state embeddings on a fixed prompt set. For LLMDNA, we largely follow the original pipeline, with slight modifications to adjust for our hardware and specific models in our

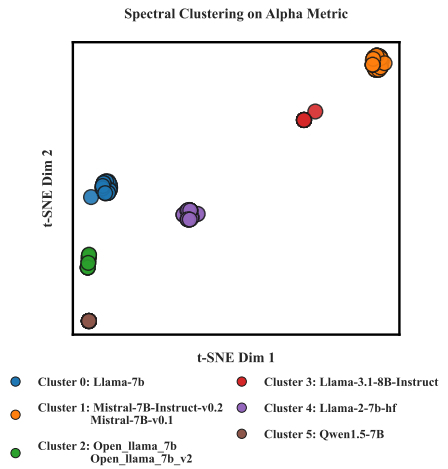

Figure 8. t-SNE 2D embedding for spectral clustering on PL_Alpha_Hill with ground truth cluster labels.

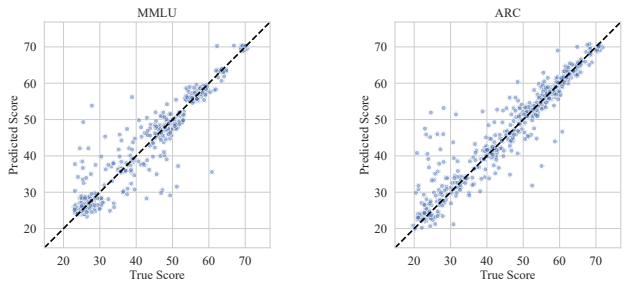

Figure 9. Predicted versus ground-truth benchmark scores on MMLU and ARC across the model pool. The dashed diagonal line denotes perfect agreement.

dataset. For our method, we compute the spectral signatures directly from model weights without any forward passes. We measure *Run Time* from the end of model loading to the end of signature/embedding extraction. One iteration of EmbedLLM embeddings takes 2,633 seconds, and one iteration of LLMDNA embeddings takes 5,393 seconds, while the Runtime of Spectral Signature is 337 seconds, significantly more time-efficient than the baselines.

7.8. Additional Generalization Tests

Model Lineage on MoE models We evaluated our spectral signature on Pangu-Pro-MoE(Tang et al., 2025), a Mixture-of-Experts model whose architecture is derived from the Qwen2.5-14B family. For the MLP modules, we averaged the metrics across experts for comparison. Table 7 shows the overall similarity between Pangu-Pro-MoE and several candidate models. Our method identifies Qwen2.5-14B, Qwen2.5-Coder-14B, and OpenMath-Nemotron-14B as the most likely source models. This result is obtained with only a minor modification, demonstrating that our method generalizes to MoE architectures with minimal adaptation.

Table 7. Similarity between Pangu-Pro-MoE and candidate source models.

Candidate model	Similarity
Qwen/Qwen2.5-14B	0.9076
nvidia/OpenMath-Nemotron-14B	0.8652
Qwen/Qwen2.5-Coder-14B	0.8354
Qwen/Qwen1.5-14B	0.3032
01-ai/Yi-1.5-9B	0.0688
meta-llama/Llama-2-13b-hf	0.0351

Similarity between models with different architectures Models derived from the same source may differ in depth. We propose an order-preserving layer matching via dynamic programming (DP) on spectral signatures. We build a cost matrix $D[i, j] = mean|Z_1[i] - Z_2[j]|$ over 7 z-score normalized module columns, then find the injective order-preserving mapping minimizing total cost via DP. We evaluate on the same model pair (Llama-3.1-8B vs. Llama-3.2-3B) to enable direct comparison with their activation-based ground truth.

Table 8 shows a 3.7× lower matching cost for the related pair. Moreover, the optimal matching skips layers 11, 17, 23, 29 of Llama-3.1-8B, exactly the correspondence recovered by (Zhu et al., 2025) via activation-based matching, validating that our data-free spectral signatures carry sufficient structural information for cross-depth layer alignment.

Table 8. Matching cost and post-matching Spearman correlation

Candidate Model	**Matching cost (↓)**	**Similarity (↑)**
Llama-3.2-3B	**0.256**	**0.893**
Qwen2.5-7B	0.953	0.771

Model Scale Generalization. As shown in Tab. 9, we evaluate additional models at 1.3B (GPT-Neo family) and 70B (Llama-3.1 family); related pairs achieve similarity $\geq$ 0.99 while unrelated pairs score 0.55–0.69, consistent with the main results.

Table 9. Similarity scores for additional model scales.

Model	**Relation**	**Similarity**
Llama-3.1-70B-Instruct	Related	0.99
OpenMath2-Llama3.1-70B	Related	0.99
Qwen2.5-72B	Unrelated	0.55
lamini-neo-1.3B	Related	1.00
lamini-cerebras-1.3B	Unrelated	0.68
Quokka-1.3B	Unrelated	0.69

Intra-family Performance Prediction. To further validate that spectral signatures capture within-family performance variation, we evaluate fine-tuned models with substantial benchmark gains over their base models (> 5 points). As shown in Tab. 10, the predicted scores track the

true scores rather than collapsing to the base-model prediction. For the Llama-13B family, our method achieves MAEs of 1.94 on TruthfulQA and 1.55 on MMLU, with significant rank correlation in 7 out of 8 settings.

Table 10. Intra-family performance prediction using statistically significant Kendall's Tau correlation.

Benchmark	Llama-2-13b	Llama-13b
MMLU	0.8347	0.0004
ARC-C	0.0272	0.0002
HellaSwag	0.0061	0.0048
TruthfulQA	0.0086	0.0091

7.9. Additional Details on Ablation Studies

Output-Invariant Reparameterization We apply scale-preserving reparameterization to consecutive linear maps while keeping the composed operator unchanged. We define a scaling factor $a = \exp(\gamma z)$ with strength γ and $z \sim \mathcal{N}(0, 1)$. The weights are updated as $Q' = aQ, K' = K/a$ (and similarly for V, O). For Permutation on Hidden Units, we generate a random permutation matrix P and construct a transformed weight set: $W' = \{PW_{gate}, PW_{up}, W_{down}P^{-1}\}$ for each layer. The similarity between reparameterized models and their origins remains 1 for Llama-2-7b, Qwen-1.5-7b, and Mistral-7b, suggesting robustness of our method. In contrast, PCS degrades substantially under both transformations.

Noise Perturbation For each weight matrix W_ℓ, we sample $Z_\ell \sim \mathcal{N}(0, I)$ with the same shape and define

$$\Delta W_\ell = \mathrm{std}(W_\ell) Z_\ell, \qquad W_\ell^{(\gamma)} = W_\ell + \gamma \Delta W_\ell, \quad (15)$$

where γ controls the noise strength. We pre-generate $\Delta W_{\ell\ell}$ once and reuse it across all γ values for controlled comparisons.