Who Built This Model? Tracing LLM Lineage via Spectral Fingerprints in Weight Space

Yiwei Chen[†,⋆] **Bingqi Shang**[†,⋆] **Sijia Liu**[†]
[†]Michigan State University [⋆]Equal contribution

GitHub Project

Abstract

Open-weight large language models (LLMs) are increasingly developed through complex, multi-stage pipelines, leading to intricate lineage relationships that reflect model origin, ownership, and evolution. Understanding these relationships is important for model provenance, governance, and supply-chain integrity. In this work, we investigate the notion of LLM "biometrics" (analogous to human biometrics) to ask whether LLMs exhibit intrinsic fingerprints in weight space alone, without access to input data, that reveal their origin and lineage. We formulate this as a lineage discrimination problem, distinguishing among independent-origin, same-series, and shared-base models. To characterize these relationships, we propose a unified *geometric fingerprinting* framework that analyzes weight matrices from two complementary perspectives: (i) spectral energy, captured by singular value distributions to encode global magnitude patterns, and (ii) subspace alignment, quantified via subspace deviations to capture directional geometry. Our analysis uncovers a clear hierarchy of structural similarity in weight space: spectral energy reliably distinguishes independently trained models and different model families, while subspace alignment enables fine-grained discrimination among closely related models, including variations in dataset scale and post-training procedures. Extensive experiments on over 110 diverse open-weight LLM pairs demonstrate that weight-space geometry provides a robust and interpretable signal for model lineage, enabling coarse-grained regime separation and fine-grained discrimination within shared-base models.

1 Introduction

The rapid growth of open-weight large language models (LLMs) has created a complex ecosystem shaped by multi-stage pipelines spanning pretraining, fine-tuning, and alignment (Touvron et al., 2023; Ouyang et al., 2022; Bai et al., 2022; Qwen, 2024; Gemma et al., 2024). This has led to a proliferation of derived models from shared base models across organizations, scales, and post-training procedures (Chung et al., 2024; Wang et al., 2023; Zhou et al., 2023; Yang et al., 2025; Chen et al., 2026a; Shang et al., 2025), forming intricate lineage relationships among models. Understanding these relationships, namely *how they are related, and how they evolve*, is critical for model governance and supply-chain integrity (Carlini et al., 2021; Pahune et al., 2025). This motivates the notion of *LLM biometrics*, which aims to identify intrinsic signatures of model identity and lineage.

Despite this need, identifying relationships between LLMs remains highly challenging. Existing approaches primarily rely on *black-box* behavioral fingerprints, such as prompt-response probing, watermarking, or output distribution analysis (Wu et al., 2025b; Kuditipudi et al., 2025; Nikolic et al., 2025; Tsai et al., 2025). These methods rely on input data, are sensitive to prompt design, and provide limited insight into intrinsic "fingerprints" or "biometrics" derived solely from LLM weights. Recent *white-box* approaches instead analyze model parameters directly, for example, using weight statistics or similarity (Zhang et al., 2024b; Zeng et al., 2025; 2024; Yoon et al., 2025). However, these methods typically reduce complex

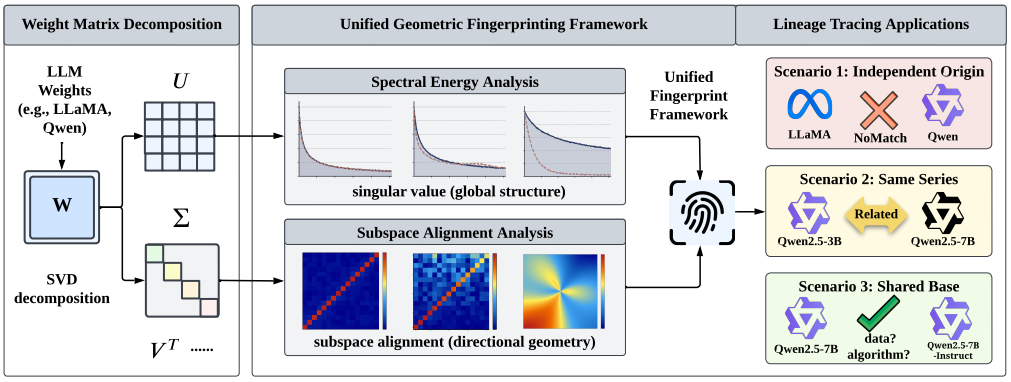


Figure 1: **Overview of the proposed geometric fingerprinting framework for LLM lineage analysis.** Given model weight matrices, we apply SVD to decompose them into two complementary representations: (i) *spectral energy* (Σ), which captures global magnitude patterns and enables coarse-grained discrimination across model families and scales, and (ii) *subspace alignment* (U), which models directional relationships via subspace alignment and supports fine-grained differentiation among closely related models. These complementary signals are unified into a single fingerprint, revealing hierarchical model relationships and enabling robust lineage tracing in weight space.

weight-space information to coarse summary statistics. In this paper, we will show that such approaches can fail to distinguish models in many lineage discrimination scenarios.

These limitations suggest that existing approaches fail to capture intrinsic structural signals that reflect model lineage. This raises a fundamental question:

***(Q)** What weight-space structures serve as reliable fingerprints for LLM lineage identification?*

In this work, we study model lineage from a weight-only perspective under a white-box setting, where model parameters are accessible while training data and metadata are not (Casper et al., 2024; Mökander, 2023). Such access is increasingly relevant given the growing availability of open-weight models and the need for transparent model auditing.

We formulate this problem as a *hierarchical lineage discrimination* task with three regimes of increasing shared information: (S1) *independent-origin*, where models are trained independently; (S2) *same-series*, where same series models share architectures and training pipelines but differ primarily in parameter scale; and (S3) *shared-base*, where models originate from a common pretrained model and diverge through post-training. These regimes serve as representative reference points rather than an exhaustive taxonomy of all possible lineage relationships. A desirable approach should therefore operate at multiple levels, separating these regimes while also characterizing intermediate relationships and distinguishing fine-grained variations within *shared-base* models (*e.g.*, differences induced by data scale and post-training), and generalizing across heterogeneous settings.

Table 1: Comparison of the applicability of existing methods—including Accurate Weight Matrix (AWM) (Zeng et al., 2025), HUman-REadable Fingerprint (HuREF) (Zeng et al., 2024), and Parameter Distribution Fingerprint (PDF) (Yoon et al., 2025)—and our approach across model lineage scenarios (S1-S3) and data-algorithm setups. Here, IO, SF, and SB denote independent-origin, same-series, and shared-base scenarios. Data and Algo. indicate the ability to capture fine-grained variations within the shared-base regime induced by dataset scale and post-training algorithms.

	S1: IO	S2: SF	S3: SB	Data.	Algo.
AWM	?	?	✓	?	?
HuREF	?	?	✓	?	?
PDF	?	?	✓	?	?
Ours	✓	✓	✓	✓	✓

To address this, we propose a unified *geometric fingerprinting framework* that analyzes weight matrices from two complementary perspectives, as **Fig. 1** shows. Model relationships exhibit both coarse-grained structure across regimes and fine-grained variations within the shared-base setting; however, existing methods fail to capture both simultaneously (as summarized in **Table 1** and will be detailed in Sec. 3). We characterize *spectral energy* of singular values to capture global magnitude patterns that separate

model families and scales, and analyze directional geometry via *subspace alignment* to reveal finer variations among closely related models. These perspectives are complementary: spectral energy distinguishes coarse differences across scenarios, while directional geometry remains discriminative within the shared-base setting. It captures variations induced by data scale, post-training methods, and their resulting geometric transformations in weight space. In summary, our **key contributions** are as follows:

① We propose a hierarchical LLM lineage discrimination problem across three regimes, independent-origin, same-series, and shared-base, with further fine-grained discrimination within the shared-base setting, enabling coarse-to-fine analysis of model relationships.

② We show that spectral energy derived from singular values captures global magnitude patterns in weight space, enabling coarse-grained separation between independent-origin and same-series models.

③ We introduce a subspace alignment analysis that captures geometric variations in weight space, enabling fine-grained discrimination among closely related shared-base models.

④ We conduct extensive experiments on over 110 open-weight LLM pairs, showing that our approach achieves both coarse-grained regime separation and fine-grained discrimination, capturing variations induced by data scale and post-training.

2 Related Works

LLM Provenance and Fingerprinting. The rapid growth of open-weight LLMs and increasingly diverse development pathways, including fine-tuning, alignment, distillation, model merging, among others, have made model *provenance* increasingly challenging (Touvron et al., 2023; Zhang et al., 2024c; Ouyang et al., 2022; Zhang et al., 2026a; Bai et al., 2022; Zhang et al., 2026b; Chen et al., 2026b; Chung et al., 2024; Wang et al., 2023; Zhang et al., 2023). While *fingerprinting* may refer to embedded identifiers or behavioral signatures (Nikolic et al., 2025; Kuditipudi et al., 2025), we use it to denote passive weight-space signatures whose pairwise geometry reveals model lineage. Following (Shao et al., 2025), existing methods fall into black-box, white-box static, and white-box forward or backward approaches. Black-box methods infer identity from model behavior, including natural output differences (Pasquini et al., 2025; Yang & Wu, 2024; Iourovitski et al., 2024; Sun et al., 2025; Bitton et al., 2025; Suzuki et al., 2025) and targeted query-response probing (Gubri et al., 2024; Jin et al., 2024; Xu et al., 2025; Tsai et al., 2025). White-box static methods analyze model weights directly: HuRef (Zeng et al., 2024) constructs algebraic attention-weight fingerprints, AWM (Zeng et al., 2025) applies CKA with layer matching, and PDF (Yoon et al., 2025) correlates layer-wise parameter statistics. Recent spectral methods derive transformation-invariant signatures from attention-weight compositions. GhostSpec (Wang et al., 2026) uses singular-value spectra, while SELF (Zhang & Zheng, 2025) combines singular values and eigenvalues. Both focus on binary verification between a designated source and a suspect model while preserving source identity under model transformations. Forward methods use inference-time probing (Zhang et al., 2024b; Chen et al., 2022; Zhang et al., 2024a), whereas backward methods exploit gradients (Wu et al., 2025a). In contrast, we study hierarchical lineage across independent-origin, same-series, and shared-base regimes, combining spectral energy and subspace alignment to capture coarse relationships and fine-grained post-training differences.

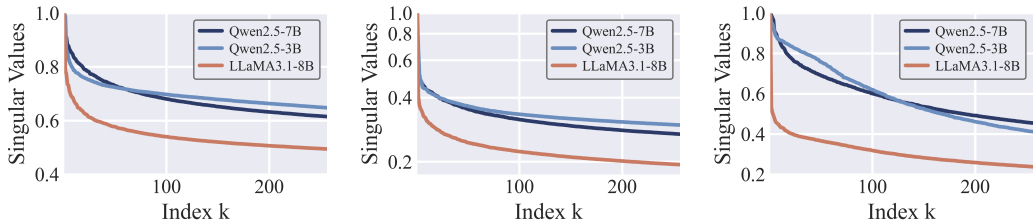
Spectral and Geometric Analysis of LLM Weights. Spectral analysis of LLM weight matrices has emerged as an effective tool for understanding representation structure and training dynamics (Martin et al., 2021; Martin & Mahoney, 2021). LLM DNA (Wu et al., 2025b) shows that projecting embeddings into a low-dimensional functional space suffices for lineage detection. Since weight matrices induce representations, identity-relevant structure should likewise be reflected in weight-space geometry. This spectral structure is largely preserved under adaptation: LoRA parameter updates yield *intruder dimensions* that exist outside the pretrained model's principal singular vectors (Shuttleworth et al., 2024; Meng et al., 2024), RLVR updates land in off-principal, low curvature subspaces (Zhu et al., 2025). Despite these task-specific deviations, the weight spaces of models with the same architecture systematically converge to a shared, low-rank subspace (Kaushik et al., 2025).

3 Tracing LLM Lineage: Motivation and Problem Statement

Weight-Space Representation of LLMs. In this work, we consider a white-box setting with access to pre-trained open-weight LLMs, but not their training or testing data. A transformer-based LLM is parameterized by θ, represented as a collection of layer-wise weight matrices (Zeng et al., 2025; Yoon et al., 2025; Zeng et al., 2024):

$$\boldsymbol{\theta} = \left\{ \mathbf{W}_c^{(l)} \;\middle|\; l \in [L],\ c \in \{Q, K, V, O, FFN_{UP}, FFN_{DOWN}, FFN_{GATE}\} \right\}. \tag{1}$$

Here, $l \in [L] = \{1, 2, \ldots, L\}$ denotes the layer index, and c indexes the component type within each layer (*e.g.*, Q, K, V, O for attention projections, and corresponding FFN components). Given two models parametrized by $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, we aim to characterize their (dis)similarity through these weight-space structures.



(a) FFN_{UP} projection (b) FFN_{GATE} projection (c) Q projection

Figure 2: Singular value distributions of middle-layer weight matrices for three component types c in (a)–(c) across models (top-256 singular values, accommodating varying matrix sizes). The x-axis denotes the singular value index k (descending), and the y-axis shows normalized singular values (scaled by the largest value). The middle layer is the $\lfloor L/2 \rfloor$-th layer of a model with L layers. Qwen2.5-3B and Qwen2.5-7B are from the same series, while LLaMA3.1-8B is from a different family.

Motivation: Spectral Signals in Weight Space and Limitations of Existing Methods. To probe structural signals in weight space, we first examine singular value spectra of weight matrices across models (**Fig. 2**). We observe consistent patterns across layers and weight types: *models from the same series* (*e.g.*, Qwen2.5-3B and Qwen2.5-7B) *exhibit highly aligned spectral decay* in their leading singular values, whereas *models from different families* (*e.g.*, LLaMA3.1-8B against others) show *clear deviations*. These differences are most pronounced in the leading singular values, which capture dominant energy and global structure. Additional visualization results are provided in Appendix A. The observations in Fig. 2 suggest that weight space encodes systematic structural signatures of model lineage, with closer models exhibiting higher spectral similarity.

Despite the clear spectral patterns in weight space, **existing methods fail to capture these distinctions**. **Fig. 3** shows model differentiation scores from two representative baselines: AWM (accurate weight-matrix fingerprinting) (Zeng et al., 2025) and PDF (parameter distribution fingerprint) (Yoon et al., 2025). We compare independent-origin pairs (●, no shared initialization or training pipeline) with same-series pairs (×, different scales sharing architecture and training pipeline) across diverse model pairs(see Appendix B). Here, AWM computes centered kernel alignment (CKA) between attention weights with Hungarian-based layer matching, while PDF measures correlation of per-layer standard deviation sequences. From the AWM perspective (y-axis), same-series pairs largely overlap with independent-origin pairs, and even with a threshold at AWM $= 10^{-2}$, many are misclassified as independent-

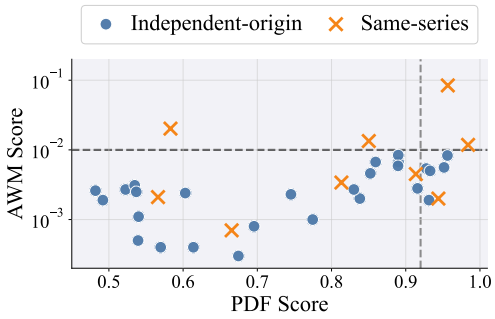


Figure 3: Normalized similarity scores in [0, 1] for model pairs under two relationship scenarios: *independent-origin* and *same-series*. Each point represents a model pair, positioned by AWM(y-axis) and PDF(x-axis) scores. Dashed lines denote the decision thresholds of AWM (horizontal) and PDF (vertical), separating regimes.

origin. From the PDF perspective (x-axis), separation is even less distinct (*e.g.*, at PDF = 0.92), with substantial overlap that makes discrimination difficult.

Proposed Hierarchical Lineage Discrimination Problem. As motivated by Fig. 3, existing similarity metrics fail to distinguish even simple lineage relationships (*e.g.*, whether models share a common origin). This raises a key question: *Can we develop a reliable lineage discrimination metric that is both sensitive and generalizable?*

Model relationships arise from shared training factors, such as initialization, data, and optimization, forming a hierarchical notion of *lineage*. As shown in **Fig. 4**, we define three regimes with increasing shared information: **independent-origin**, **same-series**, and **shared-base**. Given two models $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, we infer their relationship via a similarity function $S(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$. The model zoo of pairs that follow the hierarchy $S_{\text{independent-origin}} < S_{\text{same-series}} < S_{\text{shared-base}}$ is listed in Appendix B.

Notably, the first two regimes allow differences in architecture and scale. While such cases may appear less relevant at first glance, they serve two key purposes. First, a reliable similarity metric should generalize across this hierarchy: if it can distinguish shared-base models (*e.g.*, post-trained variants) (Ouyang et al., 2022; Lambert et al., 2024; Guo et al., 2025), it should also separate simpler coarse regimes with larger structural gaps. However, Fig. 3 shows that existing metrics fail to generalize even in these easier settings. Second, identifying model origin (*e.g.*, brand or source) across heterogeneous models is itself valuable, as models from the same source can exhibit consistent spectral signatures despite architectural differences (Fig. 2). In the literature, existing methods fail to separate coarse-grained regimes and only capture the distinctions between shared-base models with the others, as summarized in Table 1. Based on this, we adopt a coarse-to-fine perspective: global structure separates regimes, while fine-grained differences (especially within *shared-base* models) are driven by data and post-training processes.

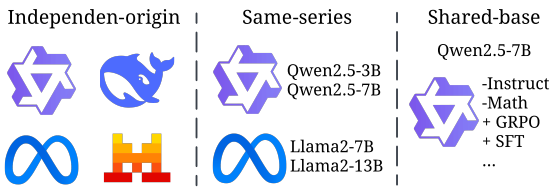


Figure 4: Model relationships form three regimes with increasing degrees of shared information (left to right). *Independent-origin:* models are trained independently without shared initialization, data, or training pipeline. *Same-series:* models from the same series share architectures and training pipelines, but differ in scale. *Shared-base:* models are derived from a common pretrained model and diverge through different post-training data and procedures (or vs. no post-training).

4 Coarse-Grained Lineage Discrimination via Spectral Energy

Spectral Trace Fingerprint. Given an LLM ($\boldsymbol{\theta}$) with layer-wise weight matrices defined in (1), let $\mathbf{W}_c^{(l)} \in \mathbb{R}^{m \times n}$ have singular values $\{\sigma_i\}_{i=1}^n$ (assuming $n \leq m$ without loss of generality). Motivated by Fig. 2, we propose using spectral energy as the key indicator to encode $\mathbf{W}_c^{(l)}$ for lineage discrimination, given by the sum of the squared singular values, $\sum_i \sigma_i^2$.

A key advantage of using spectral energy as a model fingerprint is its computational efficiency, as it can be readily obtained via the trace operation on the weights $\mathbf{W}_c^{(l)}$. That is,

$$t(\mathbf{W}_c^{(l)}) := \sqrt{\operatorname{tr}\left((\mathbf{W}_c^{(l)})^\top \mathbf{W}_c^{(l)}\right)} = \sqrt{\sum_{i=1}^n \sigma_i^2}, \tag{2}$$

where $t(\mathbf{W}_c^{(l)})$ refers to the trace-based spectral energy. Based on (2), we then construct layer-wise fingerprints for each component c of the model $\boldsymbol{\theta}$:

$$\tau_c(\boldsymbol{\theta}) := \left[t(\mathbf{W}_c^{(1)}), \ldots, t(\mathbf{W}_c^{(L)})\right]. \tag{3}$$

We refer to $\tau_c(\boldsymbol{\theta})$ as the *spectral trace fingerprint* of component type c in model $\boldsymbol{\theta}$.

Model Lineage Discrimination Metric: Trace. Given the vector-wise fingerprints in (3), we next develop a scalar-valued, easy-to-compare metric for model lineage discrimination.

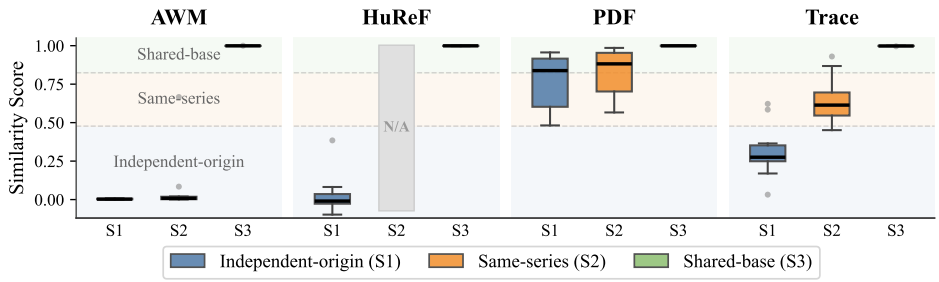


Figure 5: Distribution of Trace-based similarity scores (5) (normalized to [0,1]) for distinguishing model relationships across three scenarios: *(S1) independent-origin*, *(S2) same-series*, and *(S3) shared-base*. For comparison, distributions of baseline methods (AWM, HuReF, PDF) are also shown. The x-axis denotes ground-truth lineage scenarios, each containing corresponding model pairs, while shaded regions on the y-axis indicate desired scenario-specific score ranges, with boundaries defined by the mean trace scores. Detailed settings and results are provided in Appendix D.

This metric quantifies the relationship between two models and determines their lineage category, *i.e.*, independent-origin, same-series, or shared-base, as defined in Sec. 3.

For two models $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, we extract their trace fingerprints for each component type c, *i.e.*, $\tau_c(\boldsymbol{\theta}_a)$ and $\tau_c(\boldsymbol{\theta}_b)$ from (3). Since models may differ in depth, especially those in *independent-origin* and *same-series*, the resulting fingerprints can have different lengths. To facilitate model comparison, we align the fingerprints via linear interpolation by resampling each sequence to a common length:

$$\tilde{\tau}_c(\boldsymbol{\theta})[j] = \mathrm{Interp}\left(\tau_c(\boldsymbol{\theta}), \frac{j-1}{L_{\max}-1}\right), \quad j = 1, \ldots, L_{\max}, \tag{4}$$

where $L_{\max} = \max(L_a, L_b)$, with L_a and L_b denoting the numbers of layers in $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, respectively. Here, $\mathrm{Interp}(\cdot)$ denotes piecewise linear interpolation over the normalized layer index. For each component type, we compute the similarity between aligned trace fingerprints (*i.e.*, $\tilde{\tau}_c(\boldsymbol{\theta}_a)$ and $\tilde{\tau}_c(\boldsymbol{\theta}_b)$) using Pearson correlation, yielding our proposed similarity metric, **Trace**:

$$S_c(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) = \mathrm{Corr}\big(\tilde{\tau}_c(\boldsymbol{\theta}_a), \tilde{\tau}_c(\boldsymbol{\theta}_b)\big), \quad S(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) = \mathbb{E}_c[S_c(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)]. \tag{5}$$

Here, $S(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$ averages similarities across all component types to produce a scalar-valued model lineage discrimination metric, which serves as the Trace-oriented similarity measure for determining the lineage category to which a pair of models belongs. We refer readers to **Alg. A1** in Appendix C for an algorithm summary.

Coarse-Grained Lineage Discrimination Results. **Fig. 5** illustrates the distribution of the proposed Trace scores (5) across the three model lineage discrimination scenarios: *(S1) independent-origin*, *(S2) same-series*, and *(S3) shared-base*. Existing methods, including AWM (Zeng et al., 2025), PDF (Yoon et al., 2025), and HuReF (Zeng et al., 2024), as well as our proposed Trace (5), all assign high similarity scores to shared-base models. However, existing methods exhibit limited discriminative power in distinguishing (S1) from (S2). Specifically, AWM assigns similarly low scores to both independent-origin and same-series pairs; HuReF is not applicable to (S2) due to mismatched Transformer depths; and PDF, while improving separability, still exhibits noticeable overlap between the two scenarios. In contrast, our method (Trace) achieves a clearer hierarchical separation across the three scenarios. These results indicate that spectral trace captures intrinsic structural differences in weight space, leading to **Finding 1**.

Finding 1 (Global Structure and Hierarchy). Spectral trace fingerprints (3) and their associated similarity scores (5) capture global structure in weight space, enabling reliable coarse-grained separation and a clear hierarchical ordering of model lineage.

Motivated by the above, we focus on the most challenging setting for existing baselines: distinguishing *independent-origin* from *same-series* model pairs, which exhibit minimal structural differences and are often indistinguishable by prior methods (e.g., AWM and PDF). **Fig. 6** presents ROC curves for this task. Existing methods show limited separability: AWM even performs below random chance, while PDF improves discrimination but still exhibits noticeable overlap. In contrast, our spectral trace fingerprint (5) achieves substantially higher separability, as reflected by an improved AUC. Additional results in Appendix E show that Trace also perfectly separates *same-series* from *shared-base* pairs, indicating that the improved S1 versus S2 discrimination does not compromise shared-base detection. Together, these results demonstrate that spectral energy captures consistent structural differences across coarse-grained lineage regimes.

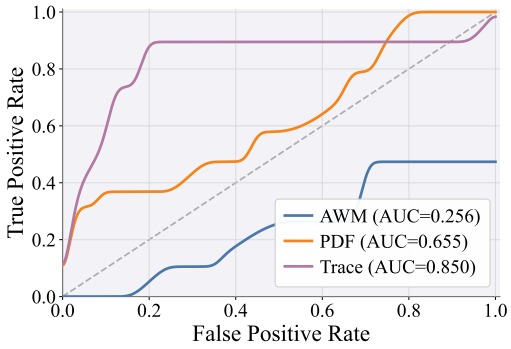


Figure 6: ROC curves for distinguishing *independent-origin* and *same-series* model pairs, experiment setting is consistent with Fig. 5. Curves are smoothed for visualization using monotonic interpolation (Fritsch & Carlson, 1980), preserving AUC values.

Based on the setup of Fig. 3, we further compare our method with the strongest baseline, PDF. **Fig. 7** plots each model pair with PDF scores on the x-axis and Trace scores on the y-axis. While PDF yields overlapping score distributions for independent-origin and same-series pairs, our method achieves clear separation along the Trace axis. Through the lens of Trace, the two studied scenarios can now be separated by a consistent margin across a horizontal decision boundary, especially when compared with AWM in Fig. 3, highlighting the improved discriminative power of our method.

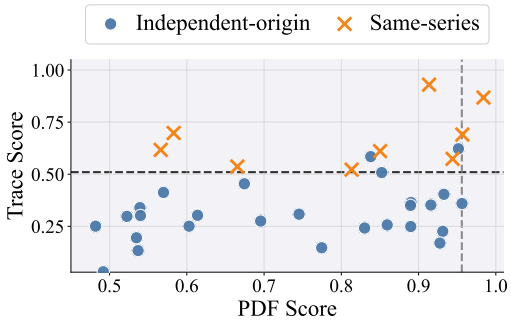


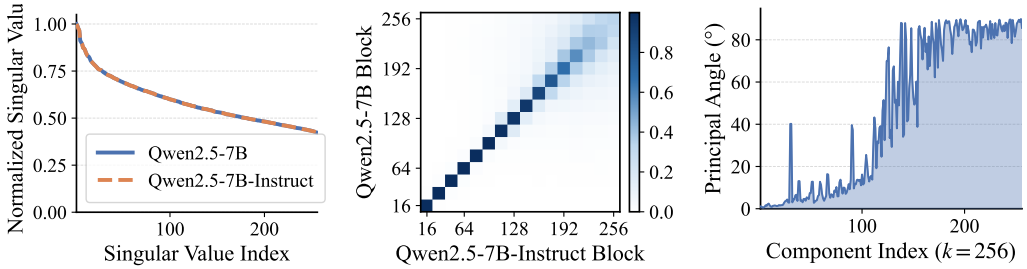
Figure 7: Comparison between Trace and PDF on distinguishing *independent-origin* and *same-series* pairs. Experiment setup stay consistent with Fig. 3

We further examine intermediate cross-series relationships between Qwen2.5 and Qwen3 models, which share a provider and aspects of model design but differ in architecture and pretraining corpus. As reported in **Table A5** in Appendix F, their mean Trace score of 0.479 lies between the independent-origin and same-series average scores. These results suggest that Trace provides a continuous relatedness signal, with the three regimes serving as representative reference points rather than exhaustive categories.

Nevertheless, Trace relies on global spectral energy and therefore provides only coarse-grained characterization. In particular, shared-base pairs receive similarity scores close to 1, making variations within this regime largely indistinguishable due to their highly aligned spectral patterns. This motivates a finer-grained analysis that captures directional differences in weight space.

5 Fine-Grained Lineage Discrimination via Subspace Alignment

Directional Signals as Complementary Structure. To examine the limitation of spectral magnitude, we analyze a representative shared-base model pair, Qwen2.5-7B and Qwen2.5-7B-Instruct (Qwen, 2024), using the attention projection Q at the 14th layer. More visualizations are offered in Appendix G. In **Fig. 8(a)**, their singular value spectra nearly overlap, indicating almost identical global structure, yielding similarity scores close to 1 under Trace (and other methods) in Fig. 5. However, spectral subspace directional differences emerge.



(a) Spectral magnitude (b) Subspace correlation (c) Subspace rotation

Figure 8: Comparison between a base model (Qwen2.5-7B) and its post-trained variant (Qwen2.5-7B-Instruct) using the attention projection Q at the middle transformer layer (14th), with top-$k = 256$ singular components. (a) Spectral magnitude: normalized singular values vs. index. (b) Subspace correlation: block-wise correlations (inner products) between groups of top-k left singular vectors from the two models. (c) Principal angles of subspace rotation: angles (in degrees) between corresponding subspaces spanned by top-k singular vectors (as used in (b)), measuring subspace rotation (Knyazev & Argentati, 2002; Zhu & Knyazev, 2012).

In **Fig. 8(b)**, while dominant subspaces remain aligned (diagonal structure), noticeable off-diagonal spread reveals local misalignment, which is measured by the correlations between groups of top-k left singular vectors from the two models across their components block. This is further reflected in **Fig. 8(c)**, where principal angles increase, especially for higher-order components, indicating progressive subspace rotation. These show that spectral magnitude in Sec. 4 captures global structure but collapses in the shared-base regime, while directional geometry provides a complementary signal that remains discriminative.

Spectral Directional Fingerprints. Unlike spectral trace fingerprints, directional analysis above requires scale-aligned weight structures for subspace comparison, and is particularly suited to model lineage analysis in the *shared-base* scenario. In what follows, we focus on model pairs with identical architectures but differing in training datasets and methods.

Given an LLM parameterized by $\boldsymbol{\theta}$ with layer-wise weight matrices $\mathbf{W}_c^{(l)} \in \mathbb{R}^{m\times n}$ as defined in (1), we perform the compact singular value decomposition (SVD):

$$\mathbf{W}_c^{(l)} = \mathbf{U}_c^{(l)} \Sigma_c^{(l)} (\mathbf{V}_c^{(l)})^\top, \tag{6}$$

where $\mathbf{U}_c^{(l)}$ and $\mathbf{V}_c^{(l)}$ denote the left and right singular vector subspace matrices, respectively, and $\boldsymbol{\Sigma}_c^{(l)}$ is a diagonal matrix with singular values as entries. From (6), we select the top-k left singular vectors $\mathbf{U}_{c,k}^{(l)}$ to define a k-dimensional subspace that captures the dominant directions (with $k = 256$ in our experiments). For each component type c, we collect subspaces across layers to form a layer-wise directional fingerprint: $\left[\mathbf{U}_{c,k}^{(1)}, \ldots, \mathbf{U}_{c,k}^{(L)}\right]$.

Rotation-based Subspace Alignment. Given two models with parameters $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, we compare their corresponding subspaces (denoted as $\mathbf{U}_{c,k}^{(l)}(\boldsymbol{\theta}_a)$ and $\mathbf{U}_{c,k}^{(l)}(\boldsymbol{\theta}_b)$ respectively) at layer l for component type c via the *cross-subspace* matrix:

$$\mathbf{C}_c^{(l)} = (\mathbf{U}_{c,k}^{(l)}(\boldsymbol{\theta}_a))^\top \mathbf{U}_{c,k}^{(l)}(\boldsymbol{\theta}_b). \tag{7}$$

Let $\{\sigma_i(\mathbf{C}_c^{(l)})\}_{i=1}^k$ denote the singular values of $\mathbf{C}_c^{(l)}$, which lie in the range $[0, 1]$ due to the orthonormality of the subspace bases (Zhu & Knyazev, 2012; Miao & Ben-Israel, 1992). These singular values correspond to the cosines of the principal angles between the two subspaces, as illustrated in Fig. 8(c). We can then measure the alignment of the subspaces $\mathbf{U}_{c,k}^{(l)}(\boldsymbol{\theta}_a)$ and $\mathbf{U}_{c,k}^{(l)}(\boldsymbol{\theta}_b)$ through the singular values of $\mathbf{C}_c^{(l)}$. For example, if the subspaces are orthogonal, then $\mathbf{C}_c^{(l)} = \mathbf{0}$. In contrast, large singular values indicate well-aligned subspaces, *i.e.*, high similarity. When all singular values equal 1, the two subspaces are identical.

Based on the above, we use the top-J smallest singular values in $\{\sigma_i(\mathbf{C}_c^{(l)})\}_{i=1}^k$ to characterize the *worst-case* subspace alignment, *i.e.*, the largest rotation between j-dimensional subspaces.

This leads to the layer-wise directional similarity score:

$$S_c^{(l)}(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b) = \frac{1}{J} \sum_{i \in \mathcal{I}_c^{(l)}} \sigma_i\left(\mathbf{C}_c^{(l)}\right) \in [0, 1], \tag{8}$$

where $\mathcal{I}_c^{(l)}$ denotes the indices of the top-J smallest singular values. We use $J = 3$ to aggregate a small tail of the least-aligned directions, preserving sensitivity to localized subspace rotations while reducing dependence on a single extreme direction. A sensitivity analysis over J is reported in **Table A6** of Appendix I.

Since $S_c^{(l)}(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$ is defined for a single component c at layer l, we obtain a scalar discrimination score $S(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$ (as in (5)) via hierarchical aggregation. Specifically, for each component c, we average the three lowest-similarity layers to obtain a component-level score, and then average across components. This focuses on layers exhibiting the most significant directional differences as measured by (8). The full procedure is summarized in **Alg. A2** , and the resulting aggregated metric based on (8) is referred to as **subspace alignment**.

Capturing Data-Scale Effects in Shared-base Models via Subspace Alignment. We consider shared-base models with varying post-training data scales, including official Instruct variants (Qwen, 2024; Yang et al., 2025; Grattafiori et al., 2024) and controlled Alpaca-SFT models (Taori et al., 2023). Our goal is to examine whether model lineage discrimination can be extended to a finer granularity, *i.e.*, to identify differences arising from distinct data-scale (post-training) in fine-tuning. In **Table 2**, existing methods produce similarity scores close to 1.0, failing to distinguish different post-training settings. In contrast, the *subspace alignment* metric yields lower and more diverse scores, revealing clear differences between models.

Two key insights emerge from Table 2, highlighting how post-training data shape directional changes. **First**, official post-training variants (*e.g.*, Instruct, Thinking) exhibit noticeably lower similarity than controlled Alpaca-SFT models, indicating that diverse, large-scale training data induce more substantial directional shifts in weight space. **Second**, within the controlled Alpaca-SFT setting, similarity decreases as the fine-tuning data scale increases from 10% to 100% (from 0.976 to 0.889), revealing a clear relationship between data scale and subspace alignment. These results demonstrate that directional similarity effectively captures data-scale effects, enabling fine-grained discrimination among shared-base models, whereas magnitude-based methods fail. This observation leads to **Finding 2**.

Finding 2 (Sensitivity to Data Scale). Subspace alignment stays sensitive to post-training data scale, capturing fine-grained subspace deviations and enabling discrimination among shared-base models that differ in fine-tuning data quantity and diversity.

Table 2: Similarity assessment for *shared-base* model pairs under different post-training settings. The upper block corresponds to official variants trained with large-scale and diverse data sources, whereas the lower block contains controlled SFT (Supervised Fine-tuned) models trained on the dataset Alpaca using different data proportions (10%, 25%, 50%, and 100%). See Appendix J for detailed experimental settings and more numerical results.

Model Pair	AWM	HuReF	PDF	Trace	Ours
Qwen2.5-7B vs Qwen2.5-7B-Instruct	0.999	0.999	1.000	0.998	0.823
Qwen2.5-14B vs Qwen2.5-14B-Instruct	0.999	1.000	1.000	0.998	0.632
Qwen3-4B vs Qwen3-4B-Instruct	0.998	0.999	0.999	0.998	0.208
Qwen3-4B vs Qwen3-4B-Thinking	0.998	0.999	0.999	0.999	0.184
Llama-3.1-8B vs Llama-3.1-8B-Instruct	0.998	0.997	0.999	0.997	0.814
Llama-3.1-8B vs Alpaca-SFT (10%)	0.999	0.998	1.000	1.000	0.976
Llama-3.1-8B vs Alpaca-SFT (25%)	0.999	0.999	0.999	0.998	0.935
Llama-3.1-8B vs Alpaca-SFT (50%)	1.000	1.000	1.000	0.998	0.903
Llama-3.1-8B vs Alpaca-SFT (100%)	1.000	1.000	1.000	0.998	0.889

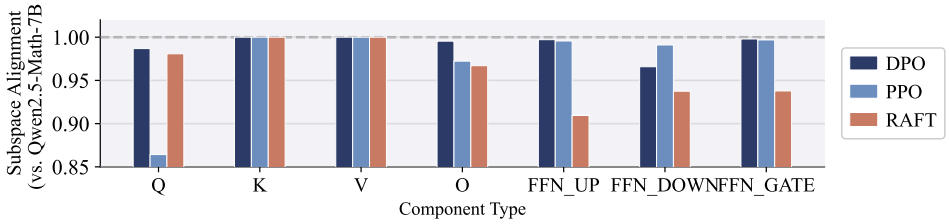


Figure 9: Per-component subspace alignment across post-training strategies with respect to the base model Qwen2.5-Math-7B. Per-component subspace alignment is defined as the average of the three smallest layer-wise similarity scores (8) across all layers, capturing worst-case directional deviation.

Algorithm-Specific Geometric Discrimination. Furthermore, we investigate how different post-training strategies affect weight-space geometry. We start from Qwen2.5-Math-7B (Yang et al., 2024) and apply different post-training approaches, including PPO (Schulman et al., 2017), DPO (Rafailov et al., 2023), and reward-ranked fine-tuning (RAFT) (Dong et al., 2023) within the Online-DPO-R1 framework (Zhang et al., 2025). All methods share the same initialization and training dataset, with aligned optimization processes. **Fig. 9** shows the sensitivity of structural similarity (measured by subspace alignment scores) to different post-training strategies across model components. The K and V projections remain close to 1.0, indicating largely preserved subspaces. In contrast, other components exhibit method-dependent deviations, reflecting how post-training perturbs weight-space alignment. RAFT yields lower similarity in FFN components, suggesting stronger perturbations, while PPO causes a noticeable drop in the Q projection. By contrast, DPO maintains consistently high similarity across components. Notably, the Q projection shows the largest variation across methods, highlighting its sensitivity to post-training dynamics and its role in capturing fine-grained geometric changes. This leads to **Finding 3**.

Finding 3 (Algorithm-Specific Subspace Sensitivity). Post-training can induce algorithm-dependent directional deviations, as measured by subspace alignment, revealing distinct geometric signatures in different model components.

Additional Experiments. Detailed numerical results on fine-grained discrimination on both data-scale and algorithmic variants are presented in Appendix K. We further evaluate robustness to quantization, unstructured magnitude pruning, model merging, and data distillation in Appendix L. Across these transformations, Trace preserves the coarse-grained shared-base lineage signal, while Subspace Alignment captures transformation severity and component-specific directional changes.

6 Conclusion

In this work, we study LLM lineage from a weight-based perspective and examine whether model identity and relationships can be inferred from intrinsic signatures. We formulate this as a hierarchical discrimination task and propose a unified geometric fingerprinting framework combining spectral and directional analysis to capture coarse-grained relationships across regimes and fine-grained variations within shared-base models. Our results show that spectral structure enables coarse-grained separation across model families, while directional geometry captures fine-grained variations induced by data scale and post-training.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation (NSF) under CISE Core Award IIS-2504263 and NSF CAREER Award IIS-2338068, the Coefficient Giving AI Safety Research Award, and the Schmidt Sciences Trustworthy AI Award.

Ethics Statement

This work proposes a passive, white-box method for analyzing LLM lineage through weight-space geometry. The method operates solely on publicly released model weights and does not involve human subjects or personal data. All models examined are open-weight and publicly available. Spectral fingerprinting is designed to support model provenance verification, intellectual-property governance, and supply-chain transparency in the rapidly expanding open-weight LLM ecosystem. As derivative models proliferate through fine-tuning, distillation, and alignment, reliable lineage attribution can help maintainers, regulators, and downstream users understand the origins of a given model and assess associated risks.

Reproducibility Statement

We take several steps to facilitate reproducibility. All algorithmic procedures, spectral trace fingerprinting in Sec. 4 and subspace alignment via principal angles in Sec. 5, are described with full mathematical detail, including similarity metrics, aggregation schemes, and hyperparameter choices both in the main paper and Appendix. Every model used in our experiments is publicly available on the HuggingFace Hub, and no proprietary data or training runs are required; fingerprint extraction operates directly on released weight files. The core computations rely on standard linear algebra routines. We report results across multiple model families, scales, and post-training algorithms to demonstrate robustness.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Yehonatan Bitton, Elad Bitton, and Shai Nisan. Detecting stylistic fingerprints of large language models. *arXiv preprint arXiv:2503.01659*, 2025.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security 21*, 2021.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. Copy, right? a testing framework for copyright protection of deep learning models. In *IEEE symposium on security and privacy (SP)*, pp. 824–841, 2022.

Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. Unlearning isn't invisible: Detecting unlearning traces in llms from model outputs. In *ICLR*, 2026a.

Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. Safety mirage: How spurious correlations undermine vlm safety fine-tuning and can be mitigated by machine unlearning. In *ICLR*, 2026b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Frederick N Fritsch and Ralph E Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.

Gemma Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

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.

Martin Gubri, Dennis Ulmer, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Trap: Targeted random adversarial prompt honeypot for black-box identification. In *Findings of ACL*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*, 2025.

Dmitri Iourovitski, Sanat Sharma, and Rakshak Talwar. Hide and seek: Fingerprinting large language models with evolutionary learning. *arXiv preprint arXiv:2408.02871*, 2024.

Heng Jin, Chaoyu Zhang, Shanghao Shi, Wenjing Lou, and Y Thomas Hou. Proflingo: A fingerprinting-based intellectual property protection scheme for large language models. In *IEEE Conference on Communications and Network Security (CNS)*, pp. 1–9, 2024.

Prakhar Kaushik, Shravan Chaudhari, Ankit Vaidya, Rama Chellappa, and Alan Yuille. The universal weight subspace hypothesis. *arXiv preprint arXiv:2512.05117*, 2025.

Andrew V Knyazev and Merico E Argentati. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040, 2002.

Rohith Kuditipudi, Jing Huang, Sally Zhu, Diyi Yang, Christopher Potts, and Percy Liang. Blackbox model provenance via palimpsestic membership inference. *arXiv preprint arXiv:2510.19796*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *JMLR*, 2021.

Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122, 2021.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *NeurIPS*, 2024.

Jianming Miao and Adi Ben-Israel. On principal angles between subspaces in rn. *Linear algebra and its applications*, 171:81–98, 1992.

Jakob Mökander. Auditing of ai: Legal, ethical and technical approaches. *Digital Society*, 2 (3):49, 2023.

Ivica Nikolic, Teodora Baluta, and Prateek Saxena. Model provenance testing for large language models. *arXiv preprint arXiv:2502.00706*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Saurabh Pahune, Zahid Akhtar, Venkatesh Mandapati, and Kamran Siddique. The importance of ai data governance in large language models. *Big Data and Cognitive Computing*, 9 (6):147, 2025.

Dario Pasquini, Evgenios M Kornaropoulos, and Giuseppe Ateniese. Llmmap: Fingerprinting for large language models. In *USENIX Security 25*, pp. 299–318, 2025.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Bingqi Shang, Yiwei Chen, Yihua Zhang, Bingquan Shen, and Sijia Liu. Forgetting to forget: Attention sink as a gateway for backdooring llm unlearning. *arXiv preprint arXiv:2510.17021*, 2025.

Shuo Shao, Yiming Li, Yu He, Hongwei Yao, Wenyuan Yang, Dacheng Tao, and Zhan Qin. Sok: Large language model copyright auditing via fingerprinting. *arXiv preprint arXiv:2508.19843*, 2025.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. *arXiv preprint arXiv:2502.12150*, 2025.

Teppei Suzuki, Ryokan Ri, and Sho Takase. Natural fingerprints of large language models. *arXiv preprint arXiv:2504.14871*, 2025.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

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 2023. *arXiv preprint arXiv:2302.13971*, 10, 2023.

Yun-Yun Tsai, Chuan Guo, Junfeng Yang, and Laurens van der Maaten. Rofl: Robust fingerprinting of language models. *arXiv preprint arXiv:2505.12682*, 2025.

Suqing Wang, Ziyang Ma, Li Xinyi, and Zuchao Li. Ghost in the transformer: Detecting model reuse with invariant spectral signatures. In *AAAI*, 2026.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *ACL*, 2023.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. Gradient-based model fingerprinting for llm similarity detection and family classification. *arXiv preprint arXiv:2506.01631*, 2025a.

Zhaomin Wu, Haodong Zhao, Ziyang Wang, Jizhou Guo, Qian Wang, and Bingsheng He. Llm dna: Tracing model evolution via functional representations. *arXiv preprint arXiv:2509.24496*, 2025b.

Zhenhua Xu, Zhebo Wang, Maike Li, Wenpeng Xing, Chunqiang Hu, Chen Zhi, and Meng Han. Rap-sm: Robust adversarial prompt via shadow models for copyright verification of large language models. *arXiv preprint arXiv:2505.06304*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhiguang Yang and Hanzhou Wu. A fingerprint for large language models. *arXiv preprint arXiv:2407.01235*, 2024.

Do-hyeon Yoon, Minsoo Chun, Thomas Allen, Hans Müller, Min Wang, and Rajesh Sharma. Intrinsic fingerprint of llms: Continue training is not all you need to steal a model! *arXiv preprint arXiv:2507.03014*, 2025.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. Huref: Human-readable fingerprint for large language models. *NeurIPS*, 2024.

Boyi Zeng, Lin Chen, Ziwei He, Xinbing Wang, and Zhouhan Lin. Awm: Accurate weight-matrix fingerprint for large language models. *arXiv preprint arXiv:2510.06738*, 2025.

Hanning Zhang, Jiarui Yao, Chenlu Ye, Wei Xiong, and Tong Zhang. Online-dpo-r1: Unlocking effective reasoning without the ppo overhead. *Notion Blog*, 2025.

Hanxiu Zhang and Yue Zheng. Self: A robust singular value and eigenvalue approach for llm fingerprinting. *arXiv preprint arXiv:2512.03620*, 2025.

Jie Zhang, Jiayuan Li, Haiqiang Fei, Lun Li, and Hongsong Zhu. Easydetector: Using linear probe to detect the provenance of large language models. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 2410–2417, 2024a.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*, 2024b.

Zhiahao Zhang, Yiwei Chen, Weizhan Zhang, Caixia Yan, Qinghua Zheng, Qi Wang, and Wangdu Chen. Tile classification based viewport prediction with multi-modal fusion transformer. In *ACM-MM*, 2023.

Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding. In *CVPR*, 2024c.

Zhihao Zhang, Abhinav Kumar, Girish Chandar Ganesan, and Xiaoming Liu. Unleashing the power of chain-of-prediction for monocular 3d object detection. In *CVPR*, 2026a.

Zhihao Zhang, Abhinav Kumar, and Xiaoming Liu. Towards intrinsic-aware monocular 3d object detection. *arXiv preprint arXiv:2603.27059*, 2026b.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL*, 2024.

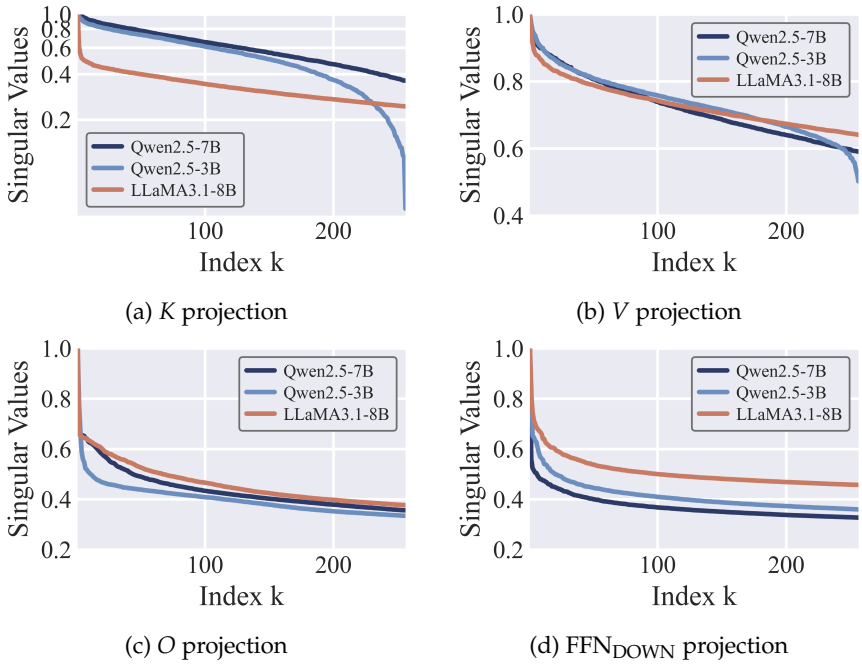
Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *NeurIPS*, 2023.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, et al. The path not taken: Rlvr provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

Peizhen Zhu and Andrew V Knyazev. Angles between subspaces and their tangents. *arXiv preprint arXiv:1209.0523*, 2012.

Appendix

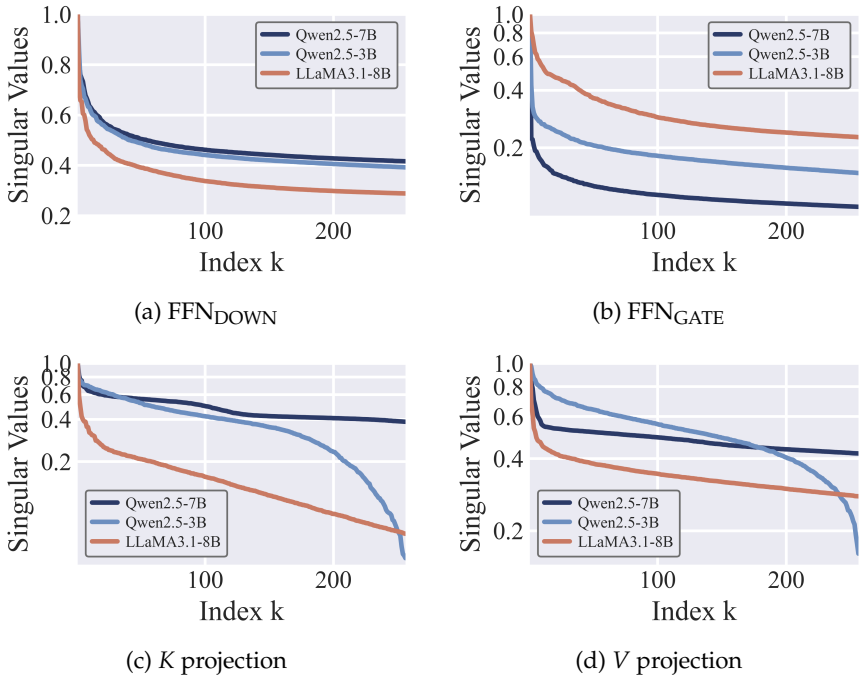
A Additional Visualizations of Spectral Signals



(a) K projection (b) V projection

(c) O projection (d) FFN_{DOWN} projection

Figure A1: Singular value distributions of middle-layer weight matrices for component types $c \in \{K, V, O, \text{FFN}_{\text{DOWN}}\}$ across models. Experiment setup is consistent with Fig. 2.



(a) FFN_{DOWN} (b) FFN_{GATE}

(c) K projection (d) V projection

Figure A2: Singular value distributions of *first-layer* weight matrices for component types $c \in \{\text{FFN}_{\text{DOWN}}, \text{FFN}_{\text{GATE}}, K, V\}$ across models. The other experiment setup is consistent with Fig. 2.

B Model Zoo of Each Regime

This section catalogs the model pairs used in our experiments, organized by the three lineage discrimination regimes defined in Sec. 3: independent-origin (Scenario 1), same-series (Scenario 2), and shared-base (Scenario 3). We reserve the same-series regime for scale variants within the same model generation, such as Qwen2.5 models of different sizes or Qwen3 models of different sizes. Cross-generation pairs between Qwen2.5 and Qwen3 are instead treated as intermediate cross-series relationships and analyzed separately in Appendix F. These pairs form the empirical basis for evaluating coarse-grained regime separation and fine-grained discrimination capabilities.

B.1 Independent-Origin Regime (Scenario 1)

Model pairs in this regime originate from different organizations or training pipelines, with no shared base model. The complete set comprises 27 model pairs across diverse model families:

Qwen cross-family pairs (7 pairs):

- Qwen/Qwen2.5-7B vs. meta-llama/Llama-2-7b-hf.
- Qwen/Qwen2.5-7B vs. meta-llama/Llama-3.1-8B.
- Qwen/Qwen2.5-7B vs. mistralai/Mistral-7B-v0.3.
- Qwen/Qwen2.5-7B vs. google/gemma-7b.
- Qwen/Qwen2.5-7B vs. 01-ai/Yi-6B.
- Qwen/Qwen2.5-7B vs. tiiuae/falcon-7b.
- Qwen/Qwen2.5-7B vs. deepseek-ai/deepseek-llm-7b-base.

LLaMA cross-family pairs (10 pairs):

- meta-llama/Llama-2-7b-hf vs. google/gemma-7b.
- meta-llama/Llama-2-7b-hf vs. mistralai/Mistral-7B-v0.3.
- meta-llama/Llama-2-7b-hf vs. 01-ai/Yi-6B.
- meta-llama/Llama-2-7b-hf vs. tiiuae/falcon-7b.
- meta-llama/Llama-2-7b-hf vs. deepseek-ai/deepseek-llm-7b-base.
- meta-llama/Llama-3.1-8B vs. mistralai/Mistral-7B-v0.3.
- meta-llama/Llama-3.1-8B vs. google/gemma-7b.
- meta-llama/Llama-3.1-8B vs. 01-ai/Yi-6B.
- meta-llama/Llama-3.1-8B vs. tiiuae/falcon-7b.
- meta-llama/Llama-3.1-8B vs. deepseek-ai/deepseek-llm-7b-base.

Other cross-family pairs (10 pairs):

- mistralai/Mistral-7B-v0.3 vs. google/gemma-7b.
- mistralai/Mistral-7B-v0.3 vs. 01-ai/Yi-6B.
- mistralai/Mistral-7B-v0.3 vs. tiiuae/falcon-7b.
- mistralai/Mistral-7B-v0.3 vs. deepseek-ai/deepseek-llm-7b-base.
- google/gemma-7b vs. 01-ai/Yi-6B.
- google/gemma-7b vs. tiiuae/falcon-7b.
- google/gemma-7b vs. deepseek-ai/deepseek-llm-7b-base.
- 01-ai/Yi-6B vs. tiiuae/falcon-7b.
- 01-ai/Yi-6B vs. deepseek-ai/deepseek-llm-7b-base.
- tiiuae/falcon-7b vs. deepseek-ai/deepseek-llm-7b-base.

These 27 pairs represent the most distinct lineage relationships, spanning organizations including Alibaba (Qwen2.5), Meta (LLaMA), Mistral AI, Google (Gemma), 01.AI (Yi), TII (Falcon), and DeepSeek. Models differ in architecture details, training data composition, vocabulary, and tokenization schemes, producing maximally divergent weight-space structures.

B.2 Same-Series Regime (Scenario 2)

Model pairs in this regime belong to the same model series and share architectures and training pipelines, but differ primarily in parameter scale. The set comprises 10 model pairs:

Qwen series (4 pairs):

- Qwen/Qwen2.5-3B vs. Qwen/Qwen2.5-7B.
- Qwen/Qwen2.5-7B vs. Qwen/Qwen2.5-14B.
- Qwen/Qwen3-4B vs. Qwen/Qwen3-8B.
- Qwen/Qwen3-8B vs. Qwen/Qwen3-14B.

Gemma series (1 pair):

- google/gemma-2b vs. google/gemma-7b.

Pythia series (2 pairs):

- EleutherAI/pythia-1b vs. EleutherAI/pythia-1.4b.
- EleutherAI/pythia-1.4b vs. EleutherAI/pythia-2.8b.

Falcon series (1 pair):

- tiiuae/falcon-7b vs. tiiuae/falcon-11b.

Mistral/Mixtral series (1 pair):

- mistralai/Mistral-7B-v0.3 vs. mistralai/Mixtral-8x7B-v0.1.

LLaMA-2 series (1 pair):

- meta-llama/Llama-2-7b-hf vs. meta-llama/Llama-2-13b-hf.

These pairs share training methodology and architectural design but exhibit scale-induced variations in weight space structure. Scale ratios range from 1.4× (Pythia-1B vs. 1.4B) to 8× (Mistral-7B vs. Mixtral-8x7B with MoE architecture).

B.3 Shared-Base Regime (Scenario 3)

Model pairs in this regime originate from a common pretrained base model and diverge through post-training procedures. We distinguish two subcategories based on fine-grained variation sources.

Data-Scale Variants. These pairs share the same base model and post-training algorithm but differ in the size of the fine-tuning dataset:

Standard instruction tuning (5 pairs):

- Qwen/Qwen2.5-7B vs. Qwen/Qwen2.5-7B-Instruct.
- Qwen/Qwen2.5-14B vs. Qwen/Qwen2.5-14B-Instruct.
- Qwen/Qwen3-4B vs. Qwen/Qwen3-4B-Instruct.
- Qwen/Qwen3-4B vs. Qwen/Qwen3-4B-Thinking.
- meta-llama/Llama-3.1-8B vs. meta-llama/Meta-Llama-3.1-8B-Instruct.

Controlled data-scale ablation (4 pairs, meta-llama/Llama-3.1-8B base):

- meta-llama/Llama-3.1-8B vs. sft_alpaca_10pct_lr1e5 (10% of Alpaca dataset).
- meta-llama/Llama-3.1-8B vs. sft_alpaca_25pct_lr1e5 (25% of Alpaca dataset).
- meta-llama/Llama-3.1-8B vs. sft_alpaca_50pct_lr1e5 (50% of Alpaca dataset).
- meta-llama/Llama-3.1-8B vs. sft_alpaca_100pct_lr1e5 (100% of Alpaca dataset).

The controlled ablation series enables systematic evaluation of how dataset scale affects weight-space geometry while holding all other factors constant.

Algorithmic Variants. These pairs share the same base model but differ in the post-training algorithm applied, isolating the effect of algorithmic choice on weight structure:

meta-llama/Llama-3.1-8B with Tulu-3 variants (3 pairs):

- meta-llama/Llama-3.1-8B vs. allenai/Llama-3.1-Tulu-3-8B.
- meta-llama/Llama-3.1-8B vs. allenai/Llama-3.1-Tulu-3-8B-SFT.
- meta-llama/Llama-3.1-8B vs. allenai/Llama-3.1-Tulu-3-8B-DPO.

Qwen/Qwen2.5-Math-7B with RLHFlow variants (5 pairs):

- Qwen/Qwen2.5-Math-7B vs. RLHFlow/Qwen2.5-7B-SFT.
- Qwen/Qwen2.5-Math-7B vs. RLHFlow/Qwen2.5-7B-DPO.
- Qwen/Qwen2.5-Math-7B vs. RLHFlow/Qwen2.5-7B-DPO-Zero.
- Qwen/Qwen2.5-Math-7B vs. RLHFlow/Qwen2.5-7B-PPO-Zero.
- Qwen/Qwen2.5-Math-7B vs. RLHFlow/Qwen2.5-7B-RAFT-Zero.

Qwen3 with reasoning-specialized variants (3 pairs) (Huan et al., 2025):

- Qwen/Qwen3-14B vs. ReasoningTransferability/UniReason-Qwen3-14B-no-think-SFT.
- Qwen/Qwen3-14B vs. ReasoningTransferability/UniReason-Qwen3-14B-think-SFT.
- Qwen/Qwen3-14B vs. ReasoningTransferability/UniReason-Qwen3-14B-RL.

These algorithmic variants span supervised fine-tuning (SFT), direct preference optimization (DPO), proximal policy optimization (PPO), and rejection sampling fine-tuning (RAFT), enabling systematic comparison of how different alignment algorithms induce geometric transformations in weight space.

Gaussian Noise Augmentation for Controlled Experiments. To systematically evaluate robustness and expand the experimental coverage, we augment the model zoo with synthetic variants created via controlled Gaussian noise perturbation. For each model in the dataset, we create a noise-perturbed copy by adding Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ to all weight matrices, where $\sigma = 0.0001$ is the noise standard deviation. This yields an additional set of model pairs with the same cardinality as the original dataset, effectively doubling the total number of pairs across all three regimes.

The noise magnitude $\sigma = 0.0001$ is chosen to be small enough to preserve model functionality (general utility) while introducing measurable weight-space perturbations.

C Spectral Magnitude-based Fingerprint Algorithm

Algorithm A1 Spectral Trace Fingerprint and Comparison

Require: Models $\mathcal{M}_a, \mathcal{M}_b$; components list $\mathcal{C}$
Ensure: Similarity score $S(\mathcal{M}_a, \mathcal{M}_b)$

for each component $c \in \mathcal{C}$ **do**
  function EXTRACTTRACE($\mathcal{M}, c$)
    $\tau_c \leftarrow \emptyset$
    for each layer l **do**
      $W \leftarrow W_c^{(l)}$
      $t \leftarrow \sqrt{\mathrm{tr}(W^\top W)}$ ▷ total spectral energy
      append t to τ_c
    end for
    return τ_c
  end function
  $\tau_c^a \leftarrow$ EXTRACTTRACE($\mathcal{M}_a, c$)
  $\tau_c^b \leftarrow$ EXTRACTTRACE($\mathcal{M}_b, c$)
  $L \leftarrow \max(|\tau_c^a|, |\tau_c^b|)$
  $\tilde{\tau}_c^a, \tilde{\tau}_c^b \leftarrow$ INTERP($\tau_c^a, \tau_c^b \to L$)
  $S_c \leftarrow \mathrm{PearsonCorr}(\tilde{\tau}_c^a, \tilde{\tau}_c^b)$
end for
$S \leftarrow \frac{1}{|\mathcal{C}|} \sum_c S_c$
return S

D Additional Settings and Results of Spectral Energy

Table A1: Similarity scores for Scenario 1 (independent-origin) before and after Gaussian perturbation ($\sigma^2 = 10^{-4}$) under baselines and our Trace-based similarity score.

Model Pair	Original				Noise			
	AWM	HuReF	PDF	Trace	AWM	HuReF	PDF	Trace
Qwen2.5-7B vs LLaMA-2-7B	0.0023	-0.066	0.7453	0.3082	0.0023	-0.066	0.745	0.3077
Qwen2.5-7B vs LLaMA-3.1-8B	0.0028	0.0092	0.9159	0.3522	0.0028	0.0092	0.916	0.352
Qwen2.5-7B vs Mistral-7B-v0.3	0.0019	0.0282	0.9313	0.2261	0.0019	0.0282	0.9312	0.227
Qwen2.5-7B vs Gemma-7B	0.0019	-0.0244	0.4919	0.0321	0.0018	-0.0244	0.492	0.1736
LLaMA-3.1-8B vs Mistral-7B-v0.3	0.0056	-0.0186	0.9516	0.6222	0.0055	-0.0186	0.9517	0.6221
LLaMA-2-7B vs Gemma-7B	0.0024	-0.0977	0.6029	0.2507	0.0024	-0.0977	0.6029	0.1412
LLaMA-2-7B vs Mistral-7B-v0.3	0.0067	0.0277	0.8594	0.2568	0.0067	0.0277	0.8594	0.2572
LLaMA-3.1-8B vs Gemma-7B	0.0027	0.0358	0.5225	0.2977	0.0027	0.0358	0.5223	0.1035
Mistral-7B-v0.3 vs Gemma-7B	0.0026	-0.0286	0.4819	0.2507	0.0026	-0.0286	0.4818	0.1174
Yi-6B vs LLaMA-2-7B	0.0067	0.0663	0.8903	0.3648	0.0067	0.0663	0.8904	0.3648
Yi-6B vs Qwen2.5-7B	0.0020	-0.0474	0.8381	0.5845	0.0019	-0.0474	0.8381	0.5838
Falcon-7B vs LLaMA-2-7B	0.0005	-0.0097	0.5393	0.3396	0.0005	-0.0097	0.5394	0.3396
Falcon-7B vs Qwen2.5-7B	0.0008	-0.0108	0.6956	0.2753	0.0008	-0.0108	0.6954	0.2753
DeepSeek-LLM-7B vs LLaMA-3.1-8B	0.0084	-0.0269	0.8899	0.2490	0.0084	-0.0269	0.890	0.2486
DeepSeek-LLM-7B vs Mistral-7B-v0.3	0.0054	0.0810	0.9276	0.1696	0.0053	0.0810	0.9276	0.1711
Qwen2.5-7B vs DeepSeek-LLM-7B	0.0027	0.0818	0.8301	0.2419	0.0027	0.0818	0.8299	0.2417
LLaMA-2-7B vs DeepSeek-LLM-7B	0.0083	0.3846	0.9562	0.3595	0.0083	0.3846	0.9562	0.3599
LLaMA-3.1-8B vs Yi-6B	0.0046	0.0178	0.8523	0.5080	0.0046	0.0178	0.8521	0.5084
LLaMA-3.1-8B vs Falcon-7B	0.0003	0.0047	0.6745	0.4542	0.0003	0.0047	0.6743	0.4542
Mistral-7B-v0.3 vs Yi-6B	0.0059	-0.0348	0.8896	0.3506	0.0059	-0.0348	0.8895	0.3501
Mistral-7B-v0.3 vs Falcon-7B	0.0004	-0.0026	0.6139	0.3030	0.0003	-0.0026	0.6137	0.3030
Gemma-7B vs Yi-6B	0.0031	0.0872	0.5348	0.1951	0.0031	0.0872	0.5348	0.0717
Gemma-7B vs Falcon-7B	0.0010	0.0034	0.7746	0.1472	0.0010	0.0034	0.7747	0.0842
Gemma-7B vs DeepSeek-LLM-7B	0.0025	-0.0321	0.5369	0.1337	0.0025	-0.0321	0.5369	0.1496
Yi-6B vs Falcon-7B	0.0004	0.0034	0.5698	0.4126	0.0004	0.0034	0.5696	0.4126
Yi-6B vs DeepSeek-LLM-7B	0.0050	-0.0136	0.9328	0.4036	0.0050	-0.0136	0.9328	0.4039
Falcon-7B vs DeepSeek-LLM-7B	0.0011	0.0004	0.5400	0.3018	0.0011	0.0004	0.5399	0.3018

Table A2: Similarity scores for Scenario 2 (same-series) before and after Gaussian perturbation ($\sigma^2 = 10^{-4}$) under baselines and our Trace-based similarity scores.

Model Pair	Original				Noise			
	AWM	HuReF	PDF	Trace	AWM	HuReF	PDF	Trace
Qwen2.5-3B vs Qwen2.5-7B	0.0020	NA	0.9440	0.5738	0.0020	NA	0.9440	0.5741
Qwen2.5-7B vs Qwen2.5-14B	0.0034	NA	0.8133	0.5220	0.0033	NA	0.8133	0.5219
Qwen3-4B vs Qwen3-8B	0.0842	NA	0.9567	0.6904	0.0842	NA	0.9567	0.6904
Qwen3-8B vs Qwen3-14B	0.0134	NA	0.8502	0.6109	0.0134	NA	0.8502	0.6109
Gemma-2B vs Gemma-7B	0.0021	NA	0.5662	0.6169	0.0021	NA	0.5662	0.0298
Pythia-1B vs Pythia-1.4B	0.0202	NA	0.5830	0.6979	0.0202	NA	0.5830	0.6979
Pythia-1.4B vs Pythia-2.8B	0.0045	NA	0.9136	0.9298	0.0046	NA	0.9136	0.6465
Falcon-7B vs Falcon-11B	0.0007	NA	0.6654	0.5368	0.0007	NA	0.6654	0.5368
Mistral-7B-v0.3 vs Mixtral-8×7B	0.6665	NA	0.9856	0.8679	0.6664	NA	0.9856	0.4265
LLaMA-2-7B vs LLaMA-2-13B	0.0118	NA	0.9840	0.4512	0.0119	NA	0.9840	0.4512

Table A3: Similarity scores for Scenario 3 (shared-base) before and after Gaussian perturbation ($\sigma^2 = 10^{-4}$) under baselines and our Trace-based similarity scores.

Model Pair	Original				Noise			
	AWM	HuReF	PDF	Trace	AWM	HuReF	PDF	Trace
Data-scale variants								
Qwen2.5-7B vs Qwen2.5-7B-Instruct	0.999	0.999	1.000	0.998	0.999	0.999	1.000	0.999
Qwen2.5-14B vs Qwen2.5-14B-Instruct	0.999	1.000	1.000	0.998	0.999	1.000	1.000	0.999
Qwen3-4B vs Qwen3-4B-Instruct	0.998	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Qwen3-4B vs Qwen3-4B-Thinking	1.000	0.999	0.998	0.999	1.000	0.999	0.998	0.999
LLaMA-3.1-8B vs LLaMA-3.1-8B-Instruct	0.998	0.997	0.999	0.997	0.999	0.998	0.999	0.998
LLaMA-3.1-8B vs Alpaca-10%	0.999	0.998	1.000	1.000	0.999	0.999	1.000	1.000
LLaMA-3.1-8B vs Alpaca-25%	0.999	0.999	0.999	0.998	0.999	0.999	0.999	0.999
LLaMA-3.1-8B vs Alpaca-50%	1.000	1.000	1.000	0.998	1.000	1.000	1.000	0.999
LLaMA-3.1-8B vs Alpaca-100%	1.000	1.000	1.000	0.998	1.000	1.000	1.000	0.999
Algorithmic variants								
LLaMA-3.1-8B vs Tulu-3-8B	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.999
LLaMA-3.1-8B vs Tulu-3-8B-SFT	0.999	0.999	1.000	0.998	0.999	0.999	1.000	0.999
LLaMA-3.1-8B vs Tulu-3-8B-DPO	0.998	0.999	0.999	0.999	0.999	0.999	0.999	1.000
Qwen2.5-Math-7B vs Qwen2.5-7B-SFT	1.000	1.000	0.999	0.999	1.000	1.000	0.999	0.999
Qwen2.5-Math-7B vs Qwen2.5-7B-DPO	0.998	0.996	0.998	0.998	0.999	0.998	0.999	0.999
Qwen2.5-Math-7B vs Qwen2.5-7B-DPO-Zero	1.000	1.000	0.999	0.999	1.000	1.000	0.999	1.000
Qwen2.5-Math-7B vs Qwen2.5-7B-PPO-Zero	0.997	0.998	0.997	0.999	0.998	0.999	0.998	1.000
Qwen2.5-Math-7B vs Qwen2.5-7B-RAFT-Zero	1.000	1.000	0.999	0.999	1.000	1.000	0.999	1.000
Qwen3-14B vs UniReason-no-think	0.998	0.999	0.998	0.999	0.999	0.999	0.999	0.999
Qwen3-14B vs UniReason-think	1.000	0.999	1.000	0.999	1.000	0.999	1.000	1.000
Qwen3-14B vs UniReason-RL	0.998	0.999	0.999	0.997	0.999	0.999	0.999	0.998

In Scenario 1 (independent-origin), models exhibit substantial structural differences in weight space, where AWM and HuReF concentrate near zero and PDF saturates at high values, while Trace produces a wide range of scores that clearly separates model families and remain stable under perturbation. In Scenario 2 (same-series), similarity increases due to shared architecture and training pipeline, with PDF showing partial saturation, whereas the Trace-based discrimination method consistent variations by model scale and preserves these differences under noise. In Scenario 3 (shared-base), all methods yield near-saturated scores close to 1 due to shared initialization; however, Trace still reveals small but consistent deviations across both data-scale and algorithmic variants, capturing subtle directional changes that remain robust under perturbation.

E Discrimination across Adjacent Lineage Regimes

We further examine discrimination across both adjacent regime boundaries. Distinguishing independent-origin (S1) from same-series (S2) models is particularly challenging because models within the same series may differ substantially in scale while retaining related spectral structure. By contrast, shared-base (S3) models inherit the same pretrained checkpoint and are therefore expected to exhibit substantially stronger similarity.

Table A4: AUC scores for distinguishing adjacent lineage regimes. S1, S2, and S3 denote independent-origin, same-series, and shared-base, respectively.

Method	S1 vs. S2	S2 vs. S3
AWM	0.256	1.000
PDF	0.655	1.000
Trace	**0.850**	1.000

As shown in Table A4, Trace substantially improves the challenging S1-versus-S2 distinction while retaining perfect separation between S2 and S3. The highest same-series Trace score is 0.930, whereas the lowest shared-base score is 0.997, leaving a clear margin between the two regimes. Thus, the improved sensitivity to family-level relationships does not compromise shared-base identification.

F Intermediate Cross-Series Relationships

The three regimes represent increasing amounts of shared training information, but they do not exhaust all possible lineage relationships. We therefore examine Qwen2.5 and Qwen3 as an intermediate cross-series case. These models share a provider and aspects of their design philosophy, while differing in architecture and pretraining corpus.

Table A5: Trace scores for cross-series Qwen2.5 and Qwen3 model pairs. Regime-level means are included as reference points.

Model Pair	Trace
Qwen2.5-7B vs. Qwen3-14B	0.310
Qwen2.5-7B vs. Qwen3-8B	0.441
Qwen2.5-14B vs. Qwen3-8B	0.576
Qwen2.5-14B vs. Qwen3-14B	0.588
Cross-series mean	**0.479**
Independent-origin mean	0.311
Same-series mean	0.650

Table A5 shows that the cross-series scores range from 0.310 to 0.588, with a mean of 0.479. This mean lies between those of the independent-origin and same-series regimes, while the variation across individual pairs reflects the heterogeneous nature of cross-series relationships. These results support interpreting Trace as a continuous relatedness signal, with the three regimes serving as reference anchors rather than rigid or exhaustive categories.

G Additional Visualizations of Directional Signals

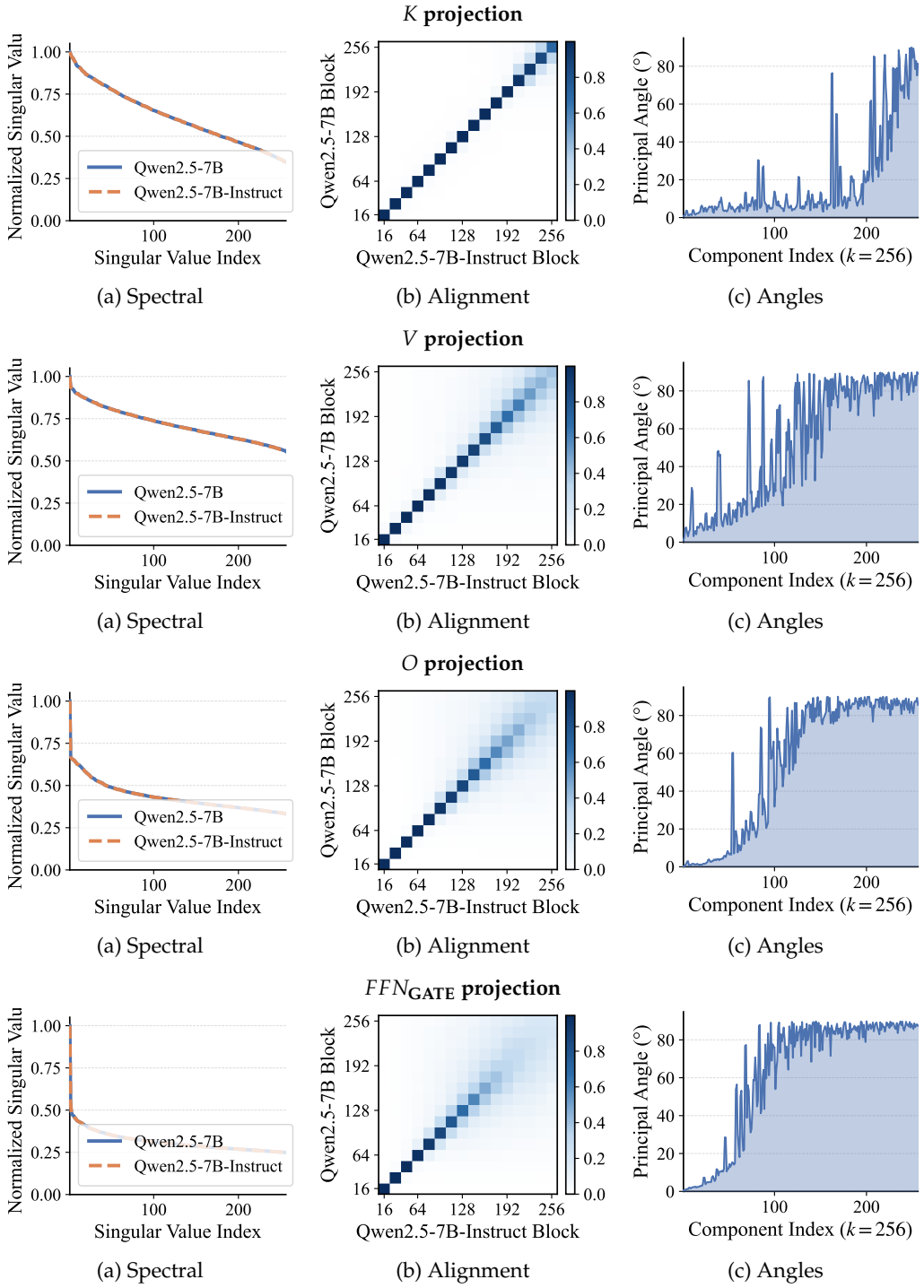


Figure A3: Comparison between a base model (Qwen2.5-7B) and its post-trained variant across multiple feature types (K, V, O, and FFN_{GATE}) at the middle transformer layer, using top-$k = 256$ singular components. Experimental setup keep consistent with Fig 8.

H Spectral Directional-based Fingerprint Algorithm

Alg. A2 summarizes the proposed Subspace Alignment algorithm. It computes layer-wise similarity from the J smallest singular values of the cross-subspace matrix, then aggregates the least-aligned layers and model components into an overall score.

Algorithm A2 Subspace Alignment Fingerprint and Comparison

Require: Models $\mathcal{M}_a, \mathcal{M}_b$; components $\mathcal{C}$; top-k; K (singular values); K_{layer}
Ensure: Similarity score $S(\mathcal{M}_a, \mathcal{M}_b)$
for each component $c \in \mathcal{C}$ **do**
  function EXTRACTSUBSPACE($\mathcal{M}, c$)
    $\mathcal{U}_c \leftarrow \emptyset$
    for each layer l **do**
      $W \leftarrow W_c^{(l)}$
      $U \leftarrow$ Top-k left singular vectors of W
      append U to $\mathcal{U}_c$
    end for
    return $\mathcal{U}_c$
  end function
  $\mathcal{U}_c^a \leftarrow$ EXTRACTSUBSPACE($\mathcal{M}_a, c$)
  $\mathcal{U}_c^b \leftarrow$ EXTRACTSUBSPACE($\mathcal{M}_b, c$)
  $\mathcal{S}_c \leftarrow \emptyset$ ▷ store layer-wise similarities
  for each layer l **do**
    $U_a \leftarrow \mathcal{U}_c^a[l], \quad U_b \leftarrow \mathcal{U}_c^b[l]$
    $C \leftarrow U_a^\top U_b$
    $\{\sigma_i\} \leftarrow \text{SingularValues}(C)$
    $\mathcal{I} \leftarrow$ indices of K smallest σ_i
    $S_c^{(l)} \leftarrow \frac{1}{K} \sum_{i \in \mathcal{I}} \sigma_i$
    append $S_c^{(l)}$ to $\mathcal{S}_c$
  end for
  $\mathcal{L}_c \leftarrow$ indices of K_{layer} smallest values in $\mathcal{S}_c$
  $S_c \leftarrow \frac{1}{K_{\text{layer}}} \sum_{l \in \mathcal{L}_c} \mathcal{S}_c[l]$
end for
$S \leftarrow \frac{1}{|\mathcal{C}|} \sum_c S_c$
return S

I Sensitivity to the Aggregation Parameter J

The parameter J in Eq. (8) determines how many of the least-aligned singular directions are averaged in the layer-wise Subspace Alignment score. As shown in Fig. 8(c), directional deviations are concentrated in a small subset of the $k = 256$ directions. A smaller J emphasizes the strongest rotations, while a larger J incorporates more well-aligned directions. Table A6 reports results for $J \in \{1, 3, 5, 10, 20\}$ across representative shared-base model pairs. The scores increase monotonically with J, while the relative differences across post-training settings remain consistent. We use $J = 3$ to retain sensitivity to localized geometric changes while reducing dependence on a single extreme direction.

Table A6: Subspace Alignment scores under different values of J for representative shared-base model pairs. DPO, PPO, and RAFT are compared with Qwen2.5-Math-7B, while Alpaca variants are compared with LLaMA-3.1-8B. The main experiments use $J = 3$.

Setting	$J = 1$	$J = 3$	$J = 5$	$J = 10$	$J = 20$
DPO	0.875	**0.957**	0.974	0.986	0.993
PPO	0.970	**0.975**	0.993	0.996	0.998
RAFT	0.933	**0.961**	0.986	0.992	0.996
Alpaca-10%	0.930	**0.976**	0.985	0.992	0.995
Alpaca-100%	0.637	**0.889**	0.921	0.959	0.979

J Additional Experimental Setups of Shared-Base Models

This section documents the controlled experimental procedures for constructing data-scale variants, which enable systematic evaluation of how fine-tuning dataset size affects weight-space geometry. All shared-base model pairs are cataloged in Appendix B.

Data-Scale Variants. To evaluate sensitivity to dataset scale, we construct controlled variants by fine-tuning the same base model on datasets of different sizes. We conduct a systematic ablation study using meta-llama/Llama-3.1-8B as the base model, creating four data-scale variants by subsampling the Alpaca instruction dataset (Taori et al., 2023):

- **10% scale**: 5,200 instruction-response pairs (10% of 52K Alpaca dataset).
- **25% scale**: 13,000 instruction-response pairs (25% of dataset).
- **50% scale**: 26,000 instruction-response pairs (50% of dataset).
- **100% scale**: 52,000 instruction-response pairs (full Alpaca dataset).

Implementation via LLaMA-Factory. All data-scale variants are fine-tuned using LLaMA-Factory (Zheng et al., 2024), a unified framework for efficient LLM training. We use the following configuration to ensure perfect experimental control:

Training hyperparameters:

- Learning rate: 1×10^{-5} (constant schedule).
- Batch size: 128 (across 4 GPUs with gradient accumulation).
- Training epochs: 3.
- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01.
- Gradient clipping: max norm 1.0.
- Precision: bfloat16 mixed precision training.

Data preparation: For each scale, we perform stratified random sampling from the full Alpaca dataset to ensure balanced representation across instruction categories (e.g., open-ended generation, closed QA, summarization). The same random seed (42) is used across all scales to ensure that smaller datasets are strict subsets of larger ones (i.e., $10\% \subset 25\% \subset 50\% \subset 100\%$). This nested sampling design enables direct interpretation of scale effects without confounding from data distribution shifts.

K Additional Experiments of Fine-Grained Lineage Discrimination

In the shared-base regime, all magnitude-based metrics (AWM, HuReF, PDF, and Trace) produce near-saturated scores close to 1, reflecting highly aligned global structure inherited from the same pretrained model. As a result, they fail to provide meaningful discrimination across different post-training variants.

In contrast, subspace alignment exhibits a significantly wider range of values, revealing clear distinctions across both data-scale and algorithmic variants. For example, data-scale variations induce gradual changes in alignment scores, while different optimization strategies (e.g., DPO, PPO, RAFT) lead to more pronounced deviations, indicating stronger geometric transformations in weight space.

Importantly, these patterns remain consistent under Gaussian perturbation, demonstrating that subspace alignment captures robust and intrinsic directional differences. These results highlight that fine-grained lineage signals are primarily encoded in subspace geometry rather than spectral magnitude.

L Robustness under Model Transformations

We further evaluate the robustness of our framework under four common model transformations in the shared-base setting. Specifically, we apply INT8 and INT4 quantization to Qwen2.5-7B, unstructured magnitude pruning with sparsity ranging from 10% to 30%,

Table A7: Similarity scores for Scenario 3 (shared-base) before and after Gaussian perturbation ($\sigma^2 = 10^{-4}$), including subspace alignment for fine-grained analysis.

Model Pair	Original					Noise				
	AWM	HuReF	PDF	Trace	Subspace	AWM	HuReF	PDF	Trace	Subspace
Data-scale variants										
Qwen2.5-7B vs Qwen2.5-7B-Instruct	0.999	0.999	1.000	0.998	0.823	0.999	0.999	1.000	0.999	0.824
Qwen2.5-14B vs Qwen2.5-14B-Instruct	0.999	1.000	1.000	0.998	0.632	0.999	1.000	1.000	0.999	0.631
Qwen3-4B vs Qwen3-4B-Instruct	0.998	0.999	0.999	0.999	0.208	0.999	0.999	0.999	0.999	0.209
Qwen3-4B vs Qwen3-4B-Thinking	1.000	0.999	0.998	0.999	0.184	1.000	0.999	0.998	0.999	0.185
LLaMA-3.1-8B vs LLaMA-3.1-8B-Instruct	0.998	0.997	0.999	0.997	0.814	0.999	0.998	0.999	0.998	0.810
LLaMA-3.1-8B vs Alpaca-10%	0.999	0.998	1.000	1.000	0.976	0.999	0.999	1.000	1.000	0.981
LLaMA-3.1-8B vs Alpaca-25%	0.999	0.999	0.999	0.998	0.935	0.999	0.999	0.999	0.999	0.930
LLaMA-3.1-8B vs Alpaca-50%	1.000	1.000	1.000	0.998	0.903	1.000	1.000	1.000	0.999	0.909
LLaMA-3.1-8B vs Alpaca-100%	1.000	1.000	1.000	0.998	0.889	1.000	1.000	1.000	0.999	0.872
Algorithmic variants										
LLaMA-3.1-8B vs Tulu-3-8B	1.000	1.000	1.000	0.999	0.787	1.000	1.000	1.000	0.999	0.789
LLaMA-3.1-8B vs Tulu-3-8B-SFT	0.999	0.999	1.000	0.998	0.783	0.999	0.999	1.000	0.999	0.783
LLaMA-3.1-8B vs Tulu-3-8B-DPO	0.998	0.999	0.999	0.999	0.783	0.999	0.999	0.999	1.000	0.780
Qwen2.5-Math-7B vs Qwen2.5-7B-SFT	1.000	1.000	0.999	0.999	0.934	1.000	1.000	0.999	0.999	0.932
Qwen2.5-Math-7B vs Qwen2.5-7B-DPO	0.998	0.996	0.998	0.998	0.957	0.999	0.998	0.999	0.999	0.963
Qwen2.5-Math-7B vs Qwen2.5-7B-DPO-Zero	1.000	1.000	0.999	0.999	0.992	1.000	1.000	0.999	1.000	0.990
Qwen2.5-Math-7B vs Qwen2.5-7B-PPO-Zero	0.997	0.998	0.997	0.999	0.975	0.998	0.999	0.998	1.000	0.972
Qwen2.5-Math-7B vs Qwen2.5-7B-RAFT-Zero	1.000	1.000	0.999	0.999	0.961	1.000	1.000	0.999	1.000	0.961
Qwen3-14B vs UniReason-no-think	0.998	0.999	0.998	0.999	0.360	0.999	0.999	0.999	0.999	0.368
Qwen3-14B vs UniReason-think	1.000	0.999	1.000	0.999	0.358	1.000	0.999	1.000	1.000	0.352
Qwen3-14B vs UniReason-RL	0.998	0.999	0.999	0.997	0.356	0.999	0.999	0.999	0.998	0.347

SLERP merging between the DPO and PPO-Zero variants of Qwen2.5-Math-7B, and data distillation by comparing DeepSeek-R1-Distill-Qwen-7B with its Qwen2.5-Math-7B base model. As shown in Table A8, Trace remains between 0.999 and 1.000 across all evaluated transformations, indicating that the coarse-grained shared-base lineage signal is consistently preserved. Subspace Alignment provides a complementary measure of transformation severity: it remains high under INT8 quantization and model merging, but decreases under INT4 quantization, stronger pruning, and data distillation. Across these settings, the K and V projections remain largely unchanged, whereas Q, O, and FFN components exhibit substantially larger directional deviations.

Table A8: Robustness under quantization, unstructured magnitude pruning, model merging, and data distillation. Trace and overall Subspace Alignment are reported together with component-wise Subspace Alignment scores.

Transformation	Setting	Trace	Sub.	K	V	Q	O	Gate	Up	Down
Quantization	INT8	1.000	0.947	1.000	1.000	0.933	0.938	0.937	0.915	0.906
Quantization	INT4	1.000	0.357	1.000	1.000	0.150	0.098	0.115	0.090	0.048
Pruning	10%	1.000	0.855	1.000	1.000	0.888	0.777	0.784	0.758	0.777
Pruning	20%	1.000	0.679	1.000	1.000	0.686	0.635	0.489	0.471	0.474
Pruning	30%	1.000	0.491	1.000	1.000	0.462	0.305	0.241	0.224	0.203
SLERP merge	vs. Base	1.000	0.965	1.000	1.000	0.977	0.986	0.934	0.983	0.875
SLERP merge	vs. DPO	1.000	0.978	1.000	1.000	0.991	0.995	0.923	0.984	0.951
SLERP merge	vs. PPO-Zero	1.000	0.972	1.000	1.000	0.985	0.996	0.925	0.987	0.910
Distillation	vs. Base	0.999	0.401	1.000	1.000	0.112	0.339	0.111	0.112	0.129