# Ghost in the Transformer: Detecting Model Reuse with Invariant Spectral Signatures

**Suqing Wang**[1*], **Ziyang Ma**[2*], **Li Xinyi**[2], **Zuchao Li**[1†]

[1]School of Artificial Intelligence, Wuhan University
[2]School of Computer Science, Wuhan University
{wangsuqing, maziyang, xyli-lucia, zcli-charlie}@whu.edu.cn

## Abstract

Large Language Models (LLMs) are widely adopted, but their high training cost leads many developers to fine-tune existing open-source models. While most adhere to open-source licenses, some falsely claim original training despite clear derivation from public models, raising pressing concerns about intellectual property protection and the need to verify model provenance. In this paper, we propose GhostSpec, a lightweight yet effective method for verifying LLM lineage without access to training data or modification of model behavior. Our approach constructs compact and robust fingerprints by applying singular value decomposition (SVD) to invariant products of internal attention weight matrices. Unlike watermarking or output-based methods, GhostSpec is fully data-free, non-invasive, and computationally efficient. Extensive experiments show it is robust to fine-tuning, pruning, expansion, and adversarial transformations, reliably tracing lineage with minimal overhead. By offering a practical solution for model verification, our method contributes to intellectual property protection and fosters a transparent, trustworthy LLM ecosystem.

**Code** — https://github.com/DX0369/GhostSpec
**Extended version** — http://arxiv.org/abs/2511.06390

## 1 Introduction

LLMs have quickly become essential for various applications in research and industry (Achiam et al. 2023; Yang et al. 2025a; Wang et al. 2025; Zhang et al. 2025; Poon et al. 2025). Due to the high cost of training LLMs from scratch (Workshop et al. 2022), many developers modify open-source LLMs via fine-tuning, continued pre-training, merging, and compression (Yang et al. 2024; Zhu et al. 2024; Tang et al. 2025; Yang et al. 2025b; Li et al. 2023; Hu et al. 2025). While most developers comply with open-source licenses, there have been instances of falsely claiming to have trained models "from scratch" when they are in fact repackaged or fine-tuned versions of public models (e.g., Llama3-V and MiniCPM-Llama3-V 2.5) (Yao et al. 2024). It is crucial to distinguish such intellectual property violations, which often break attribution requirements, from legitimate, licensed fine-tuning. This raises concerns about plagiarism and intellectual property violations, emphasizing the need for tools to verify model lineage.

To address these concerns, researchers have proposed various model identification methods (Sun et al. 2023b), which can be broadly classified into black-box and white-box approaches. Black-box methods identify models without accessing their weights, using techniques like behavioral fingerprinting and watermarking. However, these approaches are often sensitive to randomness, adversarial changes, or require intrusive pipeline modifications. In contrast, white-box methods leverage internal parameters. While representation-based techniques analyze hidden states or gradients, they depend on data access and are computationally expensive. Direct weight comparisons are simpler but fragile under fine-tuning or pruning.

We propose GhostSpec, a simple, data-free, and robust white-box method for verifying LLM lineage. GhostSpec is, by design, a white-box method targeting the open-weight model ecosystem where weights are accessible. Our key insight is that the spectral structure of weight matrices encodes intrinsic information about a model's origin, remaining stable under various modifications. Specifically, we apply singular value decomposition (SVD) to internal matrix products within the attention mechanism, including the query-key and value-output weight products. From the resulting spectra, we select the most significant singular values based on effective rank to compute similarity. To handle architectural variations in depth, such as those resulting from layer pruning or expansion, we develop the Penalty-based Optimal Spectral Alignment (POSA) algorithm that finds the best layer-wise correspondence between models. This yields a quantitative similarity score robust to differences in depth and architectural variations. Unlike black-box or representation-based white-box methods, GhostSpec is data-independent, requires no model modification, and has minimal computational cost.

**Our contributions are summarized as follows:**

- We propose GhostSpec, a lightweight white-box method for verifying LLM lineage from model weights, requiring no training data or architectural changes, offering a practical solution for provenance verification and IP protection in open-source LLMs.

*These authors contributed equally.
†Corresponding author.

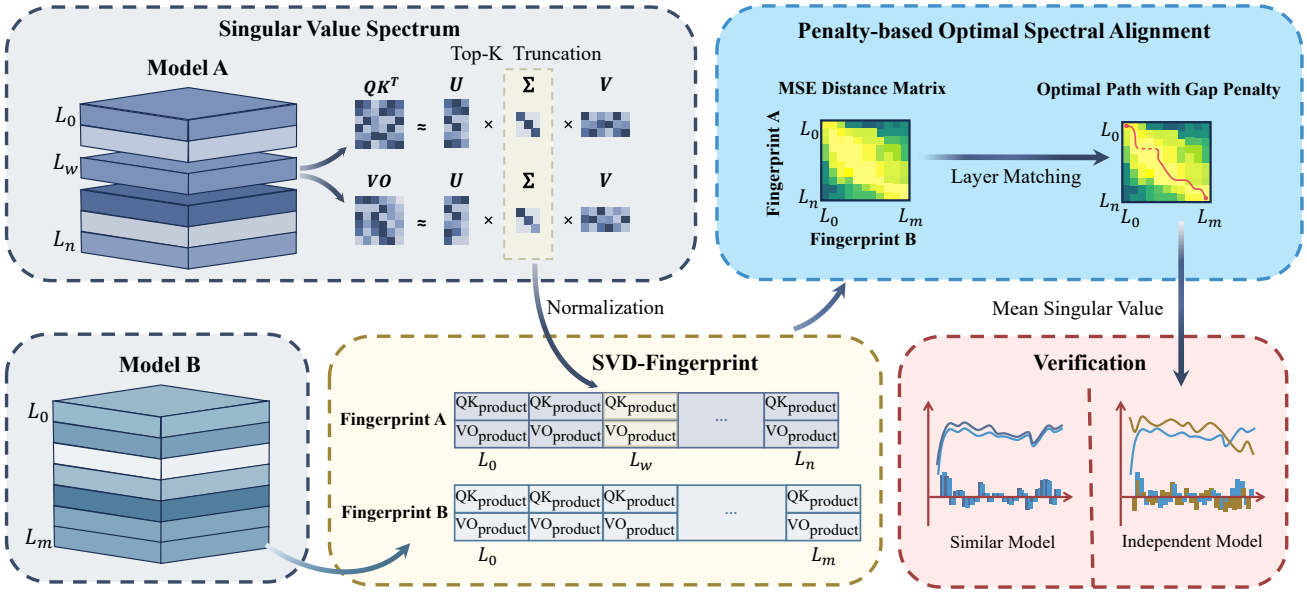


Figure 1: GhostSpec extracts singular value spectra from each layer's attention products to form spectral fingerprints. A pairwise MSE distance matrix is computed, and a penalty-based alignment algorithm matches layers across models of different depths. The final similarity score distinguishes between related and independently trained models.

- We introduce spectral fingerprints based on invariant matrix products within the attention mechanism, robust to scaling and permutation transformations, and develop the POSA algorithm for comparing models with varying depths and architectures.
- We demonstrate through experiments that GhostSpec reliably distinguishes derivative models from independently trained ones, even under challenging modifications.

## 2 Related Work

Efforts to verify model lineage can be broadly classified into black-box and white-box approaches, depending on whether internal model access is required.

### 2.1 Black-Box Identification

Black-box methods operate without access to model weights and are suitable for closed-source, API-only models. They include behavioral fingerprinting and watermarking.

**Behavioral Fingerprinting.** These passive methods identify model signatures from natural outputs. Approaches analyze stylistic or statistical patterns in generated text, or use crafted prompts to probe responses (Pasquini, Kornaropoulos, and Ateniese 2024; McGovern et al. 2024; Sam, Finzi, and Kolter 2025). Some rely on output logits or top-k probabilities to define a unique model space (Yang and Wu 2024). However, such methods are sensitive to decoding randomness and vulnerable to adversarial paraphrasing.

**Watermarking.** Watermarking embeds a detectable signal in model outputs, either via instruction tuning (Xu et al. 2024) or token-level perturbation (Kirchenbauer et al. 2023; Nagatsuka, Morishita, and Sogawa 2025). These signals can be verified statistically, but require the model creator's cooperation and can be invalidated by output editing or algorithm exposure.

### 2.2 White-Box Identification

White-box approaches leverage internal model weights or activations, typically by computing similarity between weights, intermediate representations, or gradients.

**Representation-based Fingerprinting.** These methods analyze hidden representations and generally require input data. Techniques such as CKA similarity or gradient statistics have been used to reveal shared training origins (Zhang et al. 2024; Wu, Zhao, and Wang 2025; Liang et al. 2025). While effective, these methods are computationally intensive, data-dependent, and may raise concerns regarding potential correlations with training data.

**Weight-based Fingerprinting.** This line of work focuses on static, data-free analysis of model weights. Prior methods include visualizing invariant structural features (Zeng et al. 2024) or analyzing layer-wise statistics (Yoon et al. 2025). Our proposed GhostSpec method falls into this category, capturing deeper structural information by leveraging the full singular value spectrum of invariant matrix products. The POSA algorithm is further introduced to robustly trace model ancestry under various transformations.

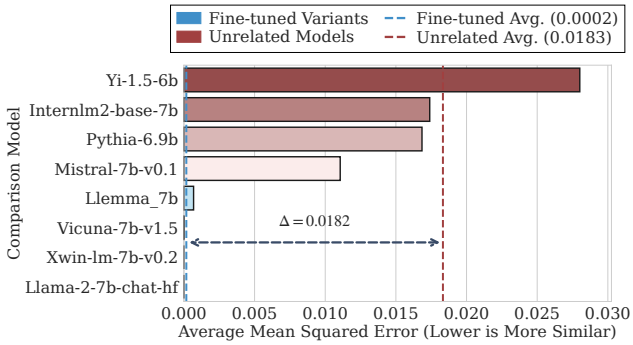


Figure 2: Average MSE of normalized singular values from Q/K/V/O projections. The spectral distance from Llama-2-7b to its fine-tuned variants (blue) is negligible, while the distance to unrelated models (red) is large, confirming the fingerprint's robustness against fine-tuning.

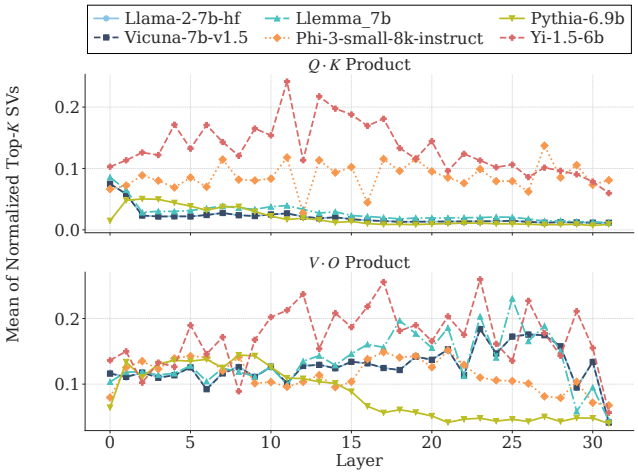


Figure 3: The layer-wise trend of the mean of normalized singular values for various models. Models with a shared lineage (e.g., Llama-2-7b-hf and its variants) exhibit highly correlated trends, while unrelated models show divergent patterns.

## 3 Preliminaries

Staats, Thamm, and Rosenow (2024) use Random Matrix Theory (RMT) to analyze the singular value spectrum of pre-trained language models. Their analysis reveals deviations from the Marchenko–Pastur distribution, especially in large singular values. These outliers correspond to dominant directions in the weight space and are strongly associated with the model's learned representations. Removing them significantly increases perplexity, showing they form a stable backbone of the model's identity.

Fine-tuning, in contrast to training from scratch, refines rather than rebuilds the model's internal structure. Staats, Thamm, and Rosenow (2024) demonstrates that fine-tuning primarily affects directions associated with small singular values. A key asymmetry is observed: pruning small singular values after fine-tuning leads to significantly greater performance degradation than pruning them before fine-tuning. This indicates that fine-tuning updates are concentrated in low-magnitude spectral components.

Together, these findings suggest spectral stability: large singular values, encoding foundational pre-trained knowledge, remain stable throughout adaptation and anchor the model's global behavior. In contrast, fine-tuning introduces targeted, low-rank modifications reflected in small singular values. We hypothesize that the large singular value spectrum of attention matrices serves as a robust, fine-tuning-invariant fingerprint.

### 3.1 Experimental Design

To empirically test this hypothesis, we designed an experiment centered on Llama-2-7b-hf. We compared this primary model against two distinct groups:

- **Fine-tuned Variants:** A group of known direct descendants of Llama-2-7b (Touvron et al. 2023), including Llama-2-7b-chat-hf, Vicuna-7b-v1.5, Llemma_7b (Azerbayev et al. 2023), and Xwin-LM-7B-V0.2 (Xwin-LM Team 2023).
- **Unrelated Models:** A control group of architecturally distinct models, including Mistral-7B-v0.1 (Jiang et al. 2023), Pythia-6.9b (Biderman et al. 2023), Yi-1.5-6B (Young et al. 2024), Internlm2-base-7b (Cai et al. 2024).

Our objective is to quantify the spectral distance between the base model and each model in these groups.

### 3.2 Quantifying Spectral Similarity

To quantify the similarity between the singular value spectra of two models $A$ and $B$, we define a compact layer-wise distance metric.

For each projection type $p \in \{q, k, v, o\}$ and each layer $i \in \{1, \ldots, L\}$, we extract the singular value vectors $\mathbf{s}_{p,A}^{(i)}$ and $\mathbf{s}_{p,B}^{(i)}$ from the corresponding weight matrices. We first truncate both vectors to their minimum effective rank $r_p^{(i)}$ and then apply min-max normalization to map them into $[0, 1]$. The spectral distance at layer $i$ and projection $p$ is then defined as the Mean Squared Error (MSE) between the normalized, truncated vectors:

$$d_p^{(i)}(A, B) = \frac{1}{r_p^{(i)}} \| \text{norm}(\text{trunc}(\mathbf{s}_{p,A}^{(i)})) - \text{norm}(\text{trunc}(\mathbf{s}_{p,B}^{(i)})) \|_2^2 . \quad (1)$$

The overall spectral distance is computed by averaging over all layers and projection types:

$$D(A, B) = \frac{1}{4L} \sum_{p \in \{q,k,v,o\}} \sum_{i=1}^{L} d_p^{(i)}(A, B). \quad (2)$$

### 3.3 Empirical Validation

As shown in Figure 2, fine-tuned models have low spectral MSE with their base models, indicating similar structures.

Algorithm 1: Penalty-based Optimal Spectral Alignment

**Input**: Distance matrix $D \in \mathbb{R}^{N \times M}$ (assume $N \leq M$), gap penalty $\rho$.
**Output**: Average distance along the optimal path.

Initialize cost matrix $C \in \mathbb{R}^{N \times M}$ and backtrack matrix $B \in \mathbb{Z}^{N \times M}$.
**for** $j = 0$ **to** $M - 1$ **do**
  $C[0, j] \leftarrow D[0, j]$ {Base case for the first layer.}
**end for**
**for** $i = 1$ **to** $N - 1$ **do**
  **for** $j = i$ **to** $M - 1$ **do**
    $k^* \leftarrow \arg\min_{i-1 \leq k < j} (C[i-1, k] + (j - k - 1)\rho)$
    $C[i, j] \leftarrow C[i-1, k^*] + (j - k^* - 1)\rho + D[i, j]$
    $B[i, j] \leftarrow k^*$
  **end for**
**end for**
$j_{\text{end}} \leftarrow \arg\min_{N-1 \leq k < M} C[N-1, k]$
Reconstruct optimal path $P$ by backtracking from $(N-1, j_{\text{end}})$ using $B$.
**return** $\frac{1}{|P|} \sum_{(i,j) \in P} D[i, j]$ {Avg. MSE on path.}

In contrast, unrelated models show much higher distances, highlighting structural dissimilarity.

These results confirm that the singular value spectrum is a stable, intrinsic property, mostly preserved through fine-tuning, motivating the design of a robust fingerprinting framework for model lineage verification under transformations.

# 4 Methodology

In practice, model weights may undergo transformation attacks such as permutation or scaling, which preserve the model's functionality while significantly altering the raw weight distribution (Zhang et al. 2024). Directly comparing the singular values of individual attention matrices $W_q$, $W_k$, $W_v$, and $W_o$ is therefore highly vulnerable to such obfuscation techniques. To address this, we propose constructing an invariant fingerprint derived from composite matrix products, which are resistant to these transformation attacks. Additionally, we introduce two complementary similarity metrics, GhostSpec-mse and GhostSpec-corr, designed to capture both fine-grained variations and macroscopic structural properties. Together, these components form a robust spectral fingerprinting methodology, as shown in Figure 1.

## 4.1 The GhostSpec Fingerprint

To quantify and compare the structural properties of Transformer-based language models, we first define an invariant spectral fingerprint that is robust to permutation and scaling transformations. For a given model $M$ with $L$ layers, we focus on the attention-related weight matrices: $W_q^{(i)}$, $W_k^{(i)}$, $W_v^{(i)}$, and $W_o^{(i)}$ from each layer $i \in \{1, \ldots, L\}$.

We define two invariant matrices per layer whose singular value spectra are resilient to functionality-preserving transformations like permutation or scaling:

$$M_{qk}^{(i)} = W_q^{(i)} (W_k^{(i)})^T \quad \text{and} \quad M_{vo}^{(i)} = W_v^{(i)} W_o^{(i)}. \tag{3}$$

These transformations alter the weight distribution without changing the model's output by modifying the weights in pairs or scaling them uniformly. Since the relative relationships between the weights are preserved, the model's core functionality and input-output behavior remain unchanged.

The fingerprint for each layer is composed of the two singular value vectors derived from these products. The layer fingerprint $\mathcal{S}_M^{(i)}$ is defined as the tuple:

$$\mathcal{S}_M^{(i)} = \left( \mathbf{s}_{qk,M}^{(i)}, \mathbf{s}_{vo,M}^{(i)} \right), \tag{4}$$

where $\mathbf{s}_{p,M}^{(i)} = \text{SVD}(M_{p,M}^{(i)})$ for $p \in \{qk, vo\}$. The complete model fingerprint $\mathcal{F}_M$ is the sequence of these layer fingerprints across all layers.

## 4.2 Similarity Metrics

Given the GhostSpec fingerprints of two models, $A$ (with $N$ layers) and $B$ (with $M$ layers), we introduce two complementary metrics: GhostSpec-mse, which provides a fine-grained, layer-by-layer comparison to measure structural correspondence, and GhostSpec-corr, a lightweight metric that captures the overall trend of spectral properties across layers. Used together, these metrics offer a comprehensive and reliable assessment of model lineage.

**Fine-grained Similarity: GhostSpec-mse.** This metric performs a detailed, layer-by-layer comparison of the singular value vectors to populate an aggregate distance matrix $D_{\text{avg}} \in \mathbb{R}^{N \times M}$. Each entry $(D_{\text{avg}})_{ij}$ represents the average spectral distance between layer $i$ of model A and layer $j$ of model B, computed as the mean of MSE over the invariant components:

$$(D_{\text{avg}})_{ij} = \frac{1}{2} \sum_{p \in \{qk, vo\}} \frac{1}{r_{p,ij}} \left\| \hat{\mathbf{s}}_{p,A}^{(i)} - \hat{\mathbf{s}}_{p,B}^{(j)} \right\|_2^2 \tag{5}$$

where:

- $p$ is the invariant product type ($qk$ or $vo$).
- $r_{p,ij} = \min(\text{eff\_rank}(\mathbf{s}_{p,A}^{(i)}), \text{eff\_rank}(\mathbf{s}_{p,B}^{(j)}))$ is the minimum of the effective ranks of the two singular value vectors being compared.
- $\hat{\mathbf{s}}$ denotes a processed singular value vector. The processing involves first truncating the original vector $\mathbf{s}$ to its top $r_{p,ij}$ values, and then applying min-max normalization to scale the result to the $[0, 1]$ range.

To handle models with different depths ($N \neq M$), we apply our POSA algorithm, as shown in Algorithm 1, to find the minimum-cost alignment path through $D_{\text{avg}}$. The raw similarity score, $d_{\text{path}}$, is the average MSE along this optimal path. Finally, we convert this distance into a normalized similarity score using an inverted Sigmoid transformation:

$$\text{Sim}_{\text{MSE}}(A, B) = 1 - \frac{1}{1 + e^{-k(d_{\text{path}} - \tau)}}, \tag{6}$$

| Primary Model: Llama-2-7b | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Method** | **Data Dep.** | **Model Fine-tuning (↑)** | | **Adversarial Transforms (↑)** | | **Unstructured Pruning (↑)** | |
| | | Vicuna-7b-v1.5 | Llemma_7b | Llama-2-7b -scaled | Llama-2-7b -permuted | Pruned-50% -Retrained | Pruned-70% -Retrained |
| **QueRE** | Data-Aware | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **Logits** | Data-Aware | 0.9767 | 0.8400 | 1.0000 | 1.0000 | 0.8567 | 0.8533 |
| **REEF** | Data-Aware | 0.9992 | 0.9979 | 1.0000 | 1.0000 | 0.9968 | 0.9948 |
| **PCS** | Data-Free | 0.9986 | 0.5052 | 0.5970 | 0.3863 | 0.9061 | 0.7829 |
| **GhostSpec-corr** | Data-Free | 0.9992 | 0.7595 | 1.0000 | 1.0000 | 0.8967 | 0.7045 |
| **GhostSpec-mse** | Data-Free | 0.9760 | 0.9532 | 0.9761 | 0.9761 | 0.9727 | 0.9653 |
| **Method** | **Data Dep.** | **Structured Pruning (↑)** | | **Merging & Expansion (↑)** | | **Unrelated Models (↓)** | |
| | | Sheared-Llama 1.3B | Sheared-Llama 2.7B | Llama2-7b-func -call-slerp | Camelidae-8x7B | Qwen2.5-7B | OPT-6.7b |
| **QueRE** | Data-Aware | 1.0000 | 1.0000 | 0.0910 | 1.0000 | 0.3410 | 1.0000 |
| **Logits** | Data-Aware | 1.0000 | 1.0000 | 1.0000 | 0.9500 | 0.9967 | 0.2200 |
| **REEF** | Data-Aware | 0.9315 | 0.9487 | 0.9996 | 0.9991 | 0.2513 | 0.2692 |
| **PCS** | Data-Free | 0.0000 | 0.0000 | 0.9993 | 0.0204 | 0.0000 | 0.0000 |
| **GhostSpec-corr** | Data-Free | 0.9398 | 0.9414 | 0.9998 | 0.9999 | 0.2940 | 0.3423 |
| **GhostSpec-mse** | Data-Free | 0.8886 | 0.9045 | 0.9760 | 0.9761 | 0.0000 | 0.5025 |
| **Primary Model: Mistral-7B** | | | | | | | |
| **Method** | **Data Dep.** | **Fine-tuning (↑)** | **Merging (↑)** | **Expansion (↑)** | **Pruning (↑)** | **Unrelated Models (↓)** | |
| | | OpenHermes-2.5 -Mistral-7B | Triunvirato-7b | Chunky-Lemon -Cookie-11B | OpenHermes-2.5 -Mistral-7B-pruned50 | Qwen2.5-7B | Yi-1.5-6B |
| **QueRE** | Data-Aware | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.3410 | 0.0819 |
| **Logits** | Data-Aware | 0.9933 | 0.9967 | 1.0000 | 0.9867 | 0.9567 | 0.2067 |
| **REEF** | Data-Aware | 0.8949 | 0.8538 | 0.8495 | 0.8596 | 0.7473 | 0.8301 |
| **PCS** | Data-Free | 0.9999 | 0.9997 | 0.8987 | 0.9979 | 0.0000 | 0.0000 |
| **GhostSpec-corr** | Data-Free | 0.9999 | 0.9997 | 0.9981 | 0.9896 | 0.2708 | 0.4304 |
| **GhostSpec-mse** | Data-Free | 0.9760 | 0.9759 | 0.9758 | 0.9753 | 0.0083 | 0.0581 |

Table 1: Comprehensive comparison of fingerprinting methods against various derivative and unrelated models, with **Llama-2-7b** and **Mistral-7B** as primary models. The table evaluates robustness to fine-tuning, architectural dissimilarity, compression, merging, expansion, and adversarial transformations. Similarity scores are color-coded based on method-specific thresholds: indicates a score above the threshold (positive classification), while indicates a score below it (negative classification).

where $\tau$ is an empirical discrimination threshold and $k$ is a steepness factor. A score approaching 1.0 indicates high similarity.

**Lightweight Similarity: GhostSpec-corr.** As illustrated in Figure 3, models with shared lineage exhibit highly correlated spectral trends, while unrelated models show divergent patterns. Based on this observation, we propose the GhostSpec-corr metric, which quantifies model similarity by capturing the overall trend of spectral properties.

First, we generate the trend sequences. For each layer $i$ and component $p$, we compute a scalar value, $\mu_{p,M}^{(i)}$, defined as the mean of the top-$K$ normalized singular values. Here, $K$ is dynamically determined based on the effective rank of each singular value spectrum, consistent with the truncation method used in GhostSpec-mse. This process produces two trend sequences for each model: $\boldsymbol{\mu}_{qk,M}$ and $\boldsymbol{\mu}_{vo,M}$.

Second, we align the sequences. Since these sequences may differ in length, direct comparison is not feasible. To address this, we apply a dynamic sequence alignment algorithm (similar to the POSA algorithm introduced previously) to match the sequence lengths. This step generates new sequences of equal length ($\boldsymbol{\mu}'_{qk,A}$, $\boldsymbol{\mu}'_{qk,B}$, etc.).

Finally, we compute the similarity. We concatenate the aligned sequences for each model and calculate the final similarity score using the distance correlation coefficient:

$$\mathrm{Sim}_{\mathrm{Corr}}(A,B) = \mathrm{dCor}\big([\boldsymbol{\mu}'_{qk,A}; \boldsymbol{\mu}'_{vo,A}], [\boldsymbol{\mu}'_{qk,B}; \boldsymbol{\mu}'_{vo,B}]\big). \tag{7}$$

This score quantifies the correlation of the models' high-level spectral evolution, providing a computationally efficient indicator of shared lineage.

## 5 Experiments

To evaluate the effectiveness and robustness of our proposed GhostSpec, we conduct a comprehensive suite of experiments designed to emulate real-world scenarios of model reuse, modification, and transformation.

### 5.1 Experimental Setup

**Dataset Construction.** We constructed a comprehensive dataset consisting of 55 model pairs, using Llama-2-7b and Mistral-7B as the primary base models. This dataset covers a wide range of transformations, including fine-tuning, com-

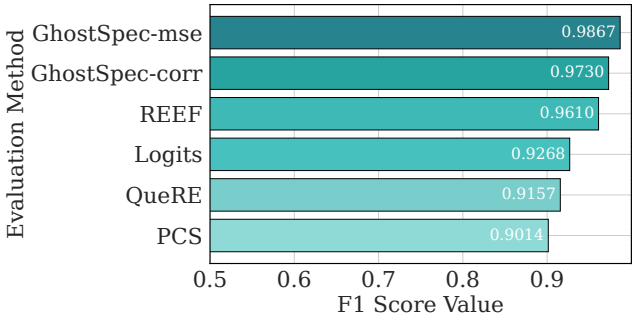


Figure 4: Maximum F1 scores for each method on our dataset. Both GhostSpec variants clearly outperform all baseline methods in accurately distinguishing related from unrelated models.

pression, merging, and expansion, with ground-truth labels indicating whether each pair is `related` or `unrelated`.

- **Fine-Tuning.** Fine-tuning is a common source of model derivation. We evaluate similarity between Llama-2-7B and several of its fine-tuned variants: Llama-2-7B-Chat-HF, Vicuna-7B-v1.5 and Llemma-7B, etc. Mistral-7B and several of its fine-tuned variants, including instruction-tuned models, DPO-optimized variants (Teknium et al. 2024).
- **Model Pruning.** For structured pruning, we consider methods that remove layers based on importance, such as ShortGPT (Men et al. 2024) and Sheared Llama (Xia et al. 2023). For unstructured pruning, we evaluate representative importance-based methods that prune weights according to their significance, including Wanda (Sun et al. 2023a) and SparseGPT (Frantar and Alistarh 2023). Ma et al. (2025) identifies critical sparsity thresholds near 35% (structured) and 60% (unstructured), beyond which model performance rapidly deteriorates. We thus evaluate sparsity levels at 30%, 50%, and 70%.
- **Model Merging.** Model merging combines weights from multiple source models to create a new model. We evaluate GhostSpec on various merging approaches, including direct weight averaging, methods that reduce parameter interference through pruning and sign conflict resolution (Yadav et al. 2023), as well as techniques that merge models by fusing weights or aligning output distributions (Goddard et al. 2024; Wan et al. 2024).
- **Model Upcycling.** Model upcycling expands model capacity by introducing new components, such as additional layers or sparsely activated expert modules, including techniques that convert dense LLMs into sparse MoEs through Parameter-Efficient Sparsity Crafting (PESC) (Wu, Zheng, and Yu 2024).
- **Permutation and Scaling Transformations.** To evaluate invariance under adversarial weight manipulations, we apply random permutations to the hidden dimensions of MLP weight matrices. For the attention layers, we apply functionality-preserving similarity transforms, which combine both random rotation and scaling, to the attention projection matrices.
- **Unrelated Models.** We computed similarity scores between our base models (Llama-2-7b and Mistral-7B) and a diverse set of independently trained LLMs, including pythia-6.9b, opt-6.7b, among others(MosaicML NLP Team 2023; Guo et al. 2024; Zhang et al. 2022).

**Evaluation Protocol.** To ensure a fair and rigorous comparison with baseline methods, we establish a standardized evaluation protocol. For each method, a model pair is classified as 'related' if its similarity score exceeds a certain threshold. We determine the optimal threshold $\tau^*$ for each method by finding the value that maximizes the F1-score against the ground truth labels (1 for related, 0 for unrelated):

$$\tau^* = \underset{\tau \in \mathcal{S}}{\operatorname{argmax}} \, \mathrm{F1}(\hat{y}_i(\tau), y_i). \tag{8}$$

where:

- $y_i$ is the true label, which is 1 for related pairs (based on ground truth) and 0 for unrelated pairs.
- $\hat{y}_i(\tau)$ is the predicted label, which is 1 if the similarity score exceeds the threshold $\tau$, and 0 otherwise.

This protocol allows us to evaluate each method at its peak performance, ensuring a fair comparison.

**Baseline Methods.** We compare GhostSpec against representative fingerprinting methods across multiple paradigms.

- **Data-Aware Baselines:** These methods require input data to generate outputs or internal representations for analysis. This category includes black-box approaches like **QueRE** (Sam, Finzi, and Kolter 2025) and **Logits** (Yang and Wu 2024), which analyze model outputs, as well as white-box methods like **REEF** (Zhang et al. 2024), which measures hidden representation similarity. A key characteristic of these methods is their reliance on data, which increases computational cost and necessitates curated datasets.
- **Data-Free Baselines:** These methods operate directly on static model weights without requiring any input data. This category includes methods like **PCS** (Zeng et al. 2024), which analyzes invariant submatrices within transformer layers.

## 5.2 Main Results

The overall classification performance of each method, measured by its maximum F1-score on our dataset, is summarized in Figure 4. Our two proposed variants, **GhostSpec-corr** (F1 = 0.9730) and **GhostSpec-mse** (F1 = 0.9867), achieve a clear lead, outperforming all data-aware and data-free baselines. Table 1 presents a selection of illustrative examples from our dataset, showing their detailed similarity scores. These scores are color-coded using the method-specific optimal thresholds, providing an intuitive visualization of the final classification results

GhostSpec demonstrated robust performance across various model transformations. It consistently produces high

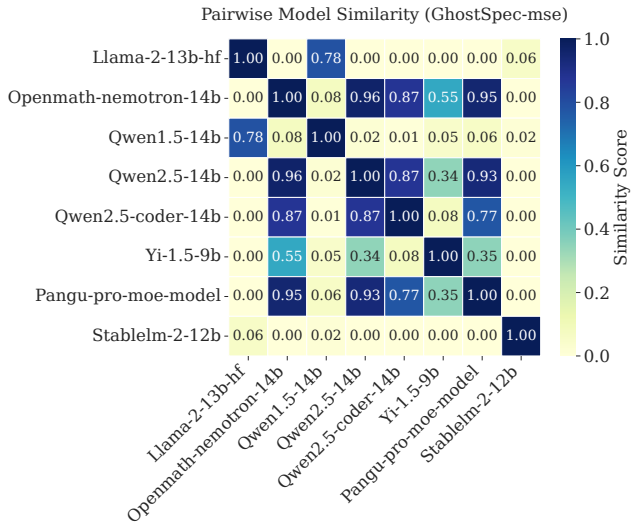


Figure 5: Pairwise structural similarity matrix of prominent open-source models computed using GhostSpec-mse. The heatmap visualizes the genealogical relationships between prominent open-source models. Higher scores indicate greater similarity.

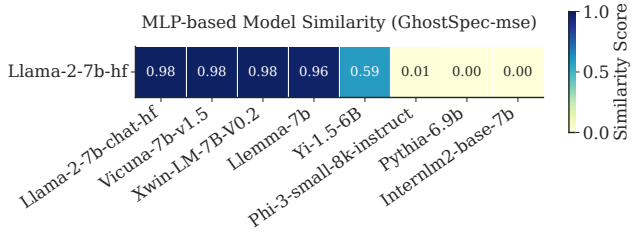


Figure 6: MLP-based spectral fingerprint similarity between Llama-2-7b and various models using GhostSpec-mse, computed from singular values of MLP up_proj and down_proj weights. Higher scores indicate greater similarity.

similarity scores for fine-tuned variants, regardless of training data or objectives. This robustness extends to structural modifications, including aggressive pruning (up to 70% sparsity), where it accurately recovers source models. It also detects strong similarity in merged and upcycled models, and remains resilient to adversarial transformations like permutation and scaling. Overall, GhostSpec reliably distinguishes true derivatives from unrelated models, showcasing strong discriminative power.

### 5.3 Further Discussion

**Resilience to Evasion Attacks.** We investigate an evasion strategy wherein an adversary fine-tunes a model with a custom loss to obscure its structural fingerprint. The objective is to maximize spectral distance from a victim model while preserving task performance.

To simulate this, we define a total loss combining task loss and a spectral divergence term:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{task}}(\theta) - \lambda \cdot D_{\text{SVD}}(\mathcal{M}_\theta, \mathcal{M}_{\text{victim}}) \tag{9}$$

where $D_{\text{SVD}}$ denotes the MSE between singular value spectra. The hyperparameter $\lambda$ balances task fidelity and spectral divergence.

We apply this method to adversarially fine-tune a Llama-3.2-1B-instruct model. Results show that it is difficult to significantly alter the singular value spectrum without hurting model performance.

These findings indicate that spectral features are intrinsically tied to model functionality, limiting the practicality of evasion via spectral manipulation.

**Case Study: A High-Profile Lineage Dispute** We apply **GhostSpec** using GhostSpec-mse similarity to investigate the recently debated lineage of Pangu-Pro-MoE by comparing it with several open-source models, including the Qwen series and Llama variants.

As shown in Figure 5, GhostSpec finds that Pangu-Pro-MoE has the highest similarity with OpenMath-Nemotron-14B (a fine-tuned variant of Qwen2.5-14B) and Qwen2.5-14B, while showing negligible similarity to unrelated models such as Yi-1.5-9B and Llama-2-13b-hf.

These results indicate a potential lineage connection between Pangu-Pro-MoE and the Qwen2.5-14B family, though further evidence is needed for confirmation.

**Analysis of MLP Module Spectra.** To examine whether spectral fingerprinting extends beyond attention layers, we analyze singular values of the MLP up-proj and down-proj weight matrices in each Transformer layer. We compute GhostSpec-mse similarity between Llama-2-7b-hf, its fine-tuned variants, and unrelated models.

As shown in Figure 6, MLP-based fingerprints effectively distinguish fine-tuned models with similarity above 0.96 from unrelated models showing near-zero similarity.

However, this approach incurs higher computational cost due to the larger size of MLP weight matrices and is less robust to common dense-to-MoE expansions where MLP layers are replaced by experts. Therefore, attention-based fingerprints remain more efficient and structurally stable for reliable lineage verification.

## 6 Conclusion

This paper tackles LLM lineage verification amid widespread reuse and potential plagiarism. We propose GhostSpec, a robust, data-independent, and efficient white-box method that constructs stable fingerprints from the singular value spectra of invariant attention matrix products. Resilient to common modifications and adversarial attacks, it employs a penalty-based optimal path alignment algorithm to handle architectural differences. Extensive experiments demonstrate GhostSpec reliably identifies model ancestry across diverse transformations, including fine-tuning, pruning, merging, and expansion. GhostSpec offers a practical, trustworthy tool to protect intellectual property and improve transparency in open-source AI ecosystems.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62306216, the Fundamental Research Funds for the Central Universities under Grant No. 2042025kf0026, and the Technology Innovation Program of Hubei Province under Grant No. 2024BAB043.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Azerbayev, Z.; Schoelkopf, H.; Paster, K.; Santos, M. D.; McAleer, S.; Jiang, A. Q.; Deng, J.; Biderman, S.; and Welleck, S. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Frantar, E.; and Alistarh, D. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, 10323–10337. PMLR.

Goddard, C.; Siriwardhana, S.; Ehghaghi, M.; Meyers, L.; Karpukhin, V.; Benedict, B.; McQuade, M.; and Solawetz, J. 2024. Arcee's MergeKit: A Toolkit for Merging Large Language Models. In Dernoncourt, F.; Preoţiuc-Pietro, D.; and Shimorina, A., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 477–485. Miami, Florida, US: Association for Computational Linguistics.

Guo, D.; Zhu, Q.; Yang, D.; Xie, Z.; Dong, K.; Zhang, W.; Chen, G.; Bi, X.; Wu, Y.; Li, Y.; et al. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming–The Rise of Code Intelligence. *arXiv preprint arXiv:2401.14196*.

Hu, J.; Li, J.; Pan, Z.; Chen, C.; Li, Z.; Wang, P.; and Zhang, L. 2025. SongSong: A Time Phonograph for Chinese SongCi Music from Thousand of Years Away. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(25), 26229–26237.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; and Goldstein, T. 2023. A watermark for large language models. In *International Conference on Machine Learning*, 17061–17084. PMLR.

Li, Q.; Li, Z.; Cai, X.; Du, B.; and Zhao, H. 2023. Enhancing visually-rich document understanding via layout structure modeling. In *Proceedings of the 31st ACM international conference on multimedia*, 4513–4523.

Liang, H.; Zheng, Y.; Li, Y.; Zhang, Y.; and Liang, S. 2025. Origin Tracer: A Method for Detecting LoRA Fine-Tuning Origins in LLMs. *arXiv preprint arXiv:2505.19466*.

Ma, Z.; Li, Z.; Zhang, L.; Xia, G.-S.; Du, B.; Zhang, L.; and Tao, D. 2025. Model Hemorrhage and the Robustness Limits of Large Language Models.

McGovern, H.; Stureborg, R.; Suhara, Y.; and Alikaniotis, D. 2024. Your large language models are leaving fingerprints. *arXiv preprint arXiv:2405.14057*.

Men, X.; Xu, M.; Zhang, Q.; Wang, B.; Lin, H.; Lu, Y.; Han, X.; and Chen, W. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*.

MosaicML NLP Team. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

Nagatsuka, K.; Morishita, T.; and Sogawa, Y. 2025. A Nested Watermark for Large Language Models. *arXiv preprint arXiv:2506.17308*.

Pasquini, D.; Kornaropoulos, E. M.; and Ateniese, G. 2024. Llmmap: Fingerprinting for large language models. *arXiv preprint arXiv:2407.15847*.

Poon, M.; Dai, X.; Liu, X.; Kong, F.; Lui, J.; and Zuo, J. 2025. Online Multi-LLM Selection via Contextual Bandits under Unstructured Context Evolution. *arXiv preprint arXiv:2506.17670*.

Sam, D.; Finzi, M.; and Kolter, J. Z. 2025. Predicting the performance of black-box llms through self-queries. *arXiv preprint arXiv:2501.01558*.

Staats, M.; Thamm, M.; and Rosenow, B. 2024. Small Singular Values Matter: A Random Matrix Analysis of Transformer Models. *arXiv preprint arXiv:2410.17770*.

Sun, M.; Liu, Z.; Bair, A.; and Kolter, J. Z. 2023a. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Sun, Y.; Liu, T.; Hu, P.; Liao, Q.; Fu, S.; Yu, N.; Guo, D.; Liu, Y.; and Liu, L. 2023b. Deep intellectual property protection: A survey. *arXiv preprint arXiv:2304.14613*.

Tang, Z.; Ma, Z.; Wang, S.; Li, Z.; Zhang, L.; Zhao, H.; Li, Y.; and Wang, Q. 2025. CoViPAL: Layer-wise Contextualized Visual Token Pruning for Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 20701–20714.

Teknium; theemozilla; karan4d; and huemin art. 2024. Nous Hermes 2 Mistral 7B DPO. https://huggingface.co/NousResearch/Nous-Hermes-2-Mistral-7B-DPO.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wan, F.; Huang, X.; Cai, D.; Quan, X.; Bi, W.; and Shi, S. 2024. Knowledge fusion of large language models. *arXiv preprint arXiv:2401.10491*.

Wang, S.; Li, Z.; Luohe, S.; Du, B.; Zhao, H.; Li, Y.; and Wang, Q. 2025. From Parameters to Performance: A Data-Driven Study on LLM Structure and Development. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 26095–26112.

Workshop, B.; Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Wu, H.; Zheng, H.; and Yu, B. 2024. Parameter-Efficient Sparsity Crafting from Dense to Mixture-of-Experts for Instruction Tuning on General Tasks. *arXiv preprint arXiv:2401.02731*.

Wu, Z.; Zhao, Y.; and Wang, H. 2025. Gradient-Based Model Fingerprinting for LLM Similarity Detection and Family Classification. *arXiv preprint arXiv:2506.01631*.

Xia, M.; Gao, T.; Zeng, Z.; and Chen, D. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.

Xu, J.; Wang, F.; Ma, M. D.; Koh, P. W.; Xiao, C.; and Chen, M. 2024. Instructional fingerprinting of large language models. *arXiv preprint arXiv:2401.12255*.

Xwin-LM Team. 2023. Xwin-LM. https://github.com/Xwin-LM/Xwin-LM. Accessed: 2023-09.

Yadav, P.; Tam, D.; Choshen, L.; Raffel, C. A.; and Bansal, M. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36: 7093–7115.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, E.; Shen, L.; Guo, G.; Wang, X.; Cao, X.; Zhang, J.; and Tao, D. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Yang, H.; Yao, Y.; Li, Z.; Qi, B.; Guoming, L.; and Zhao, H. 2025b. XQuant: Achieving Ultra-Low Bit KV Cache Quantization with Cross-Layer Compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 9796–9811.

Yang, Z.; and Wu, H. 2024. A fingerprint for large language models. *arXiv preprint arXiv:2407.01235*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yoon, D.-h.; Chun, M.; Allen, T.; Müller, H.; Wang, M.; and Sharma, R. 2025. Intrinsic Fingerprint of LLMs: Continue Training is NOT All You Need to Steal A Model! *arXiv preprint arXiv:2507.03014*.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Wang, G.; Li, H.; Zhu, J.; Chen, J.; et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zeng, B.; Wang, L.; Hu, Y.; Xu, Y.; Zhou, C.; Wang, X.; Yu, Y.; and Lin, Z. 2024. Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, 37: 126332–126362.

Zhang, J.; Liu, D.; Qian, C.; Zhang, L.; Liu, Y.; Qiao, Y.; and Shao, J. 2024. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Z.; He, L.; Li, Z.; Zhang, L.; Zhao, H.; and Du, B. 2025. Segment First or Comprehend First? Explore the Limit of Unsupervised Word Segmentation with Large Language Models. *arXiv preprint arXiv:2505.19631*.

Zhu, X.; Li, J.; Liu, Y.; Ma, C.; and Wang, W. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12: 1556–1577.