PROVABLE MODEL PROVENANCE SET FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid proliferation of customized large language models has intensified concerns over unauthorized model reuse, highlighting the need for reliable model provenance techniques for model auditing and tracking. However, existing work largely infers provenance relationships through heuristic fingerprint-matching rules and lack of provably valid error control, leaving the credibility of provenance claims unverified. In this work, we formalize model provenance as a set-valued inference problem and propose the *Model Provenance Set* (MPS), a novel principled method to adaptively construct a provenance set with rigorous coverage guarantee, i.e., covering the source models at a specified confidence level. Specifically, we introduce a novel equivalence test that yields a valid p-value under the null hypothesis that all candidates are equally close to the target model, and develop a test-and-exclusion procedure to sequentially select the most probable source candidate until the null is rejected. By establishing the sequential structure of the test, MPS constructs a subset of candidates with a provable coverage guarantee. Extensive experiments demonstrate that MPS can effectively identify all source models with the intended coverage while strictly avoiding the inclusion of unrelated models.

1 INTRODUCTION

Large language models(LLMs) are increasingly customized by fine-tuning or other adaptions, intensifying concerns over unauthorized model reuse (Tie et al., 2025; Shi et al., 2025; Zhang et al., 2025a; Yao et al., 2024a). For example, a foundation model may undergo successive rounds of fine-tuning and redistribution across multiple parties without proper licenses (Intelligence, 2024; Yao et al., 2024b; Yoon et al., 2025). Such practices highlight the growing importance of intellectual property (IP) protection across both open-source and proprietary ecosystems (Zhang & Zheng, 2025; Park et al., 2026; Nasery et al., 2025). Recently, model provenance techniques have emerged as an effective tool for identifying the source models of a target model by computing a pairwise similarity (Nikolic et al., 2025; Hu et al., 2025; Zhang et al., 2025b). Such techniques provide traceable provenance evidence, enabling mitigation of associated security, privacy, and compliance risks.

However, existing approaches largely infer the provenance relationship by heuristic rules such as predefined thresholds (Hu et al., 2025; Zhang & Zheng, 2025; Ren et al., 2025) or trained classifiers (Wu et al., 2025b; Yoon et al., 2025; Nasery et al., 2025). These methods typically lack formal statistical guarantees and remain vulnerable to misidentification. Recent work formulates model provenance as a hypothesis-testing problem, determining whether two models exhibit significant similarity relative to reference models (Nikolic et al., 2025). Nevertheless, this framework focuses on pairwise provenance determination and does not provide joint guarantees over multiple candidate source models. It thus is insufficient, since a target model may undergo successive customizations and consequently inherit provenance from multiple sources. This motivates us to develop a principled method for constructing a provenance set with rigorous coverage guarantees.

In this paper, we formalize the model provenance problem as a set-valued inference with formal error control, i.e., given a pool of candidate models, our goal is to construct a source set that covers all source models at a desired confidence level while minimizing false provenance. For this problem, we propose *Model Provenance Set* (MPS), a principled method that leverages the relative distances between candidate models and the target for provenance inference. Specifically, we develop a centered distance-based statistical test to check the source existence among candidates against the

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

null hypothesis, i.e., all candidates are equally close to the target model. We then apply a test-and-exclusion procedure to sequentially eliminates the most similar candidate until the null is accepted. At each testing step, MPS yields a valid p-value with provable control over the global false negative rate, enabling adaptive construction of the source set without error accumulation. Since MPS only requires sufficiently informative observable scores, it is agnostic to the specific choice of model similarity scores, enabling seamless integration with diverse fingerprinting techniques.

To evaluate the validity of MPS, we conduct extensive experiments on a real-world benchmark of 562 LLMs from Hugging Face (Wolf et al., 2019), spanning model sizes from 135M to 14B and across various candidate pool settings. Our results demonstrate that MPS reliably identifies all source models while excluding unrelated ones. For example, with two ground-truth source models among 50 candidates, MPS achieves a coverage rate of 96% with an average predicted set size of 2.1 at error tolerance level $\alpha = 0.05$ (Figure 2), indicating its effective control of misidentification. We further show that MPS outperforms empirical baselines on model attribution tasks and can serve as a quantitative tool for model non-infringement assessment, highlighting its applicability to *model IP protection* and *auditing*. Our code and data are available at https://anonymous.4open.science/r/MPS-CA7D.

We summarize our contributions as follows:

1. We formulate model provenance problem as model set-valued inference with formal error control: identifying a provenance set that includes all source models of the target model with at least a predefined confidence level while keeping the set as small as possible.
2. We propose **MPS** to adaptively construct a concise provenance set with rigorous coverage guarantees through a test-and-exclusion procedure. Our method provides provable asymptotic validity while remaining agnostic to the scoring strategy.
3. Extensive experiments demonstrate that MPS can consistently identify all source models with guaranteed probability while keeping the selected set compact, and enables reliable provenance analysis in diverse model attribution and auditing tasks.

2 Problem Formulation

In this section, we first introduce some necessary notations and definitions for model provenance (Section 2.1) and then formulate the model provenance problem as set-valued inference with a rigorous statistical guarantee (Section 2.2). Such rigor is crucial to ground reliable provenance claims, yet largely overlooked in prior literature (see Appendix A for a detailed review).

2.1 Definition of provenance set

In this work, we consider a general model provenance setting with multiple candidates. This setting is natural in practice, as provenance verification typically involves comparing a target model against a collection of existing models, while its provenance may involve more than one source. For example, a model may undergo successive customizations, such as general instruction following, domain specialization, or task-specific alignment (Shang et al., 2026). This motivates reasoning about the complete *set* of source models rather than a single source. We now formally define the set.

Let g denote a target model and $\mathcal{M} = \{f_1, \ldots, f_M\}$ a set of candidate models. We call f_i a *source model* of g if f_i lies in the derivation chain of g. We denote this relation by $f_i \xrightarrow{k} g$, where k denotes the number of derivation steps from f_i to g, referred to as the *lineage depth*. For example, given the derivation chain $f_i \to f_i' \to g$, the lineage depth from f_i to g is $k = 2$. Source models may occur at the same lineage depth, e.g., f_i and f_j merge to derive g. We then define the source set as follows.

Definition 2.1 (Source set)**.** *The **source set** of g with respect to $\mathcal{M}$ is defined as*

$$\mathcal{M}^* = \left\{ f_i \in \mathcal{M} : f_i \xrightarrow{k} g \text{ for some } k \geq 1 \right\}. \tag{1}$$

It contains all source models of g in $\mathcal{M}$, regardless of their lineage depths, and may be empty.

While Definition 2.1 characterizes the ideal provenance set, it is not directly obtainable in practice since model parameter transformations are generally irreversible and cannot be simply reversed to

recover the derivation chain. A statistically identifiable form is therefore needed. For this purpose, prior works primarily devise a pairwise distance score to detect the existence of provenance relationship between two model (Nikolic et al., 2025; Wu et al., 2025b; Shang et al., 2026). Specifically, let $\theta(x)$ denote the characteristic extracted from output logits or internal representations of model θ on prompt x, and $\mathcal{L}(\cdot,\cdot)$ a dissimilarity function. Given N prompts $\mathcal{X} = \{x_1, \ldots, x_N\}$, the sample-wise distance between f_i and g is

$$L_{i,t} = \mathcal{L}(g(x_t),\ f_i(x_t)), \tag{2}$$

with expected distance $\mu_i = \mathbb{E}_{t\sim\mathcal{Q}}[L_{i,t}]$, where a smaller μ_i indicates stronger signal of a provenance relationship. Existing methods typically use a fixed threshold to make a binary provenance decision (Zhang et al., 2025b; Ren et al., 2025). Yet multiple source models may exhibit varying degrees of similarity to the target, making the fixed threshold inadequate for identifying multiple sources. We therefore introduce the following assumption to characterize a statistically detectable source set.

Assumption 2.1 (Provenance separability). *Let $\mathcal{U} = \mathcal{M} \setminus \mathcal{M}^*$ and $\mathcal{M}_\mathcal{R} = \mathcal{R} \cup \mathcal{U}$ for any $\mathcal{R} \subseteq \mathcal{M}^*$. Define the studentized centered distance $\gamma_i(\mathcal{M}_\mathcal{R}) = \theta_i(\mathcal{M}_\mathcal{R})/\sigma_i(\mathcal{M}_\mathcal{R})$, where $\theta_i(\mathcal{M}_\mathcal{R}) = \mu_i - |\mathcal{M}_\mathcal{R}|^{-1}\sum_{f_j\in\mathcal{M}_\mathcal{R}}\mu_j$ and $\sigma_i^2(\mathcal{M}_\mathcal{R}) = \mathrm{Var}(L_{i,t} - |\mathcal{M}_\mathcal{R}|^{-1}\sum_{f_j\in\mathcal{M}_\mathcal{R}} L_{j,t})$. Assume there exists $\Delta_N > 0$ with $\sqrt{N}\Delta_N \to \infty$ such that in every active set $\mathcal{M}_\mathcal{R}$, there exists $f_\mathcal{R} \in \mathcal{R}$ satisfying*

$$\gamma_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R}) \le -\Delta_N \quad \textit{and} \quad \gamma_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R}) \le \min_{f_u\in\mathcal{U}} \gamma_u(\mathcal{M}_\mathcal{R}) - \Delta_N. \tag{3}$$

Assumption 2.1 generally ensures that true sources are distinguishable from unrelated candidates by a margin Δ_N, which may vary arbitrarily across models. Under a sufficient condition that source models are strictly closer to g than any unrelated candidate with constant-order separability signal $\Delta_N = O(1)$, the provenance set admits a clean population-level characterization in distances,

$$\mathcal{M}^* = \left\{ f_i \in \mathcal{M} : \mu_i < \min_{f_u\in\mathcal{U}} \mu_u \right\}, \tag{4}$$

i.e., source models are exactly those strictly closer to g than any unrelated candidate (Wu et al., 2025b; Nikolic et al., 2025). This makes $\mathcal{M}^*$ statistically accessible, providing the foundation for tractable set-valued inference for provenance task. The formal proof of the sufficient condition and the characterization of $\mathcal{M}^*$ are provided in Appendix B.1 and Appendix B.2, respectively.

2.2 Provenance Set Inference Problem

Given the definition of $\mathcal{M}^*$ in Eq. (4), we formulate provenance set inference with statistical guaranty:

$$\min_{\hat{\mathcal{M}}\subseteq\mathcal{M}} |\hat{\mathcal{M}}| \quad \text{subject to} \quad \Pr(\mathcal{M}^* \subseteq \hat{\mathcal{M}}) \ge 1-\alpha, \tag{5}$$

where $\alpha \in (0,1)$ is a user-defined significance level and $1-\alpha$ is the confidence level. Here, **coverage** refers to the event $\mathcal{M}^* \subseteq \hat{\mathcal{M}}$, and the constraint controls the false negative risk at level α.

Our goal is to infer a complete yet compact provenance set $\hat{\mathcal{M}} \subseteq \mathcal{M}$ that contains all true source models with high probability while minimizing unnecessary candidates. This objective is motivated by practical provenance verification, where the inferred set serves as a shortlist for downstream inspection. For example, in high-stakes IP auditing and compliance enforcement, provenance techniques may be used to screen a large candidate pool for a target model g, producing a compact set of candidate sources for further verification or legal assessment (Nikolic et al., 2025; Zhang et al., 2025b). The set should retain all true sources to avoid missed provenance relationships, while remaining compact to reduce downstream verification cost.

However, identifying such a set is non-trivial in practice. Existing methods typically focus on pairwise provenance determination. Although one can naively test each pair (f_i, g) and aggregate all positive results into the output set, this pairwise strategy generally does not satisfy the desired coverage constraint (see Section 5). A detailed review of prior work is provided in Appendix A. In this work, we aim to provide provable coverage guarantees by jointly testing the entire candidate set rather than performing independent pairwise comparisons, as described in the next section.

3 METHODOLOGY

This section introduces MPS, which comprises equivalence testing and exclusion operation. We first devise a test based on relative distances (Sec. 3.1) and formulate the hypothesis test for determining whether all candidates are unrelated to the target (Sec. 3.2). We then develop a sequential test-and-exclusion algorithm to adaptively identify potential source models (Sec. 3.3). The overall framework is shown in Figure 1; pseudocode and complexity analysis are deferred to Appendix C.

3.1 PROVENANCE-AWARE TEST STATISTICS

Since we usually have no a priori knowledge of which candidates are unrelated to g, we exploit the relative similarities among all candidates without relying on any external reference. Specifically, as source models are generally rarer than unrelated ones in practice (Shang et al., 2026), with sufficient candidates in $\mathcal{M}$, they are expected to exhibit significantly higher similarity to g than the ensemble average, which is dominated by unrelated models. This motivates a deviation-based test statistic.

Formally, let $L \in \mathbb{R}^{N\times|\mathcal{M}|}$ denote the distance matrix comprising the sample-wise distance values of all $f_i \in \mathcal{M}$ with the target model g. For any model f_i in the current candidate set $\mathcal{M}$, we compute the relative distances $d_{ij,t} = L_{i,t} - L_{j,t}$ and its average on the query space $\bar{d}_{ij} = \frac{1}{N}\sum_{t=1}^{N} d_{ij,t}$. The deviation of model f_i from ensemble-average distances to the target g is defined as $\bar{d}_{i\cdot} = \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}} \bar{d}_{ij}$. Note that $\bar{d}_{i\cdot} \in \mathbb{R}$ can be seen from the identity $\bar{d}_{i\cdot} = (\bar{L}_i - \bar{L}_\cdot)$, where $\bar{L}_i \equiv N^{-1}\sum_{t=1}^{N} L_{i,t}$ and $\bar{L}_\cdot \equiv M^{-1}\sum_{i=1}^{M} \bar{L}_i$. This leads to the studentized t-statistic:

$$t_i = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\mathrm{var}}(\bar{d}_{i\cdot})}}, \tag{6}$$

where $\widehat{\mathrm{var}}(\bar{d}_{i\cdot}) = \frac{1}{N}\widehat{\mathrm{var}}(d_{i\cdot,t})$ and $\widehat{\mathrm{var}}(d_{i\cdot,t}) = \frac{1}{N-1}\sum_{t=1}^{N}(d_{i\cdot,t} - \bar{d}_{i\cdot})^2$ is the standard sample variance of the observations $\{d_{i,t}\}_{t=1}^{N}$. Studentization ensures that the statistic is scale-invariant and comparable between candidates with different magnitudes of fingerprint similarity, which is critical when multiple provenance models exhibit heterogeneous similarity strengths to g.

Our purpose is to detect whether some candidate is significantly closer to the target model g than the others. We then formulate the null hypothesis in terms of their expected distances:

$$\mathrm{H}_{0,\mathcal{M}} : \mu_i = \mu_j \quad \forall f_i, f_j \in \mathcal{M}. \tag{7}$$

where $\mu_i = \mathbb{E}_{t\sim\mathcal{X}}[L_{i,t}]$. Since a smaller u_i indicates that model f_i is closer to g than the ensemble average, this motivates us to focus on the minimum deviation and define the observed statistic as

$$T_{\min,\mathcal{M}} = \min_{f_i\in\mathcal{M}} t_i. \tag{8}$$

Intuitively, under $\mathrm{H}_{0,\mathcal{M}}$, all candidates are uniformly close to g and the t-statistics $\{t_i\}$ are asymptotically centered at zero. Thus, a sufficiently small value of $T_{\min,\mathcal{M}}$ (i.e., a large negative deviation from the average level) indicates that at least one candidate is significantly closer to g than the ensemble average, suggesting the presence of a source model.

3.2 PROVENANCE TESTING USING t-STATISTICS

To assess the statistical significance of the observed $T_{\min,\mathcal{M}}$, we compute its p-value under the null by constructing a null distribution from the scores of unrelated models. Under H_0, the candidate distances are statistically indistinguishable across candidates, suggesting that these distances are exchangeable. We formalize the conditions supporting this exchangeability in the following assumption.

Assumption 3.1 (Invariance under the null)**.** *For any $\dot{\mathcal{M}} = \{f_1, \ldots, f_m\} \subseteq \mathcal{M}$, we assume that under $\mathrm{H}_{0,\dot{\mathcal{M}}}$, for any x_t and any bijection $\pi : [m] \to [m]$ (i.e., a permutation of the model ordering),*

$$(\dot{L}_{1,t}, \ldots, \dot{L}_{m,t}) \overset{d}{=} (\dot{L}_{\pi(1),t}, \ldots, \dot{L}_{\pi(m),t}),$$

i.e., the corresponding distance matrix $\dot{L}$ is distributionally invariant to model permutations.

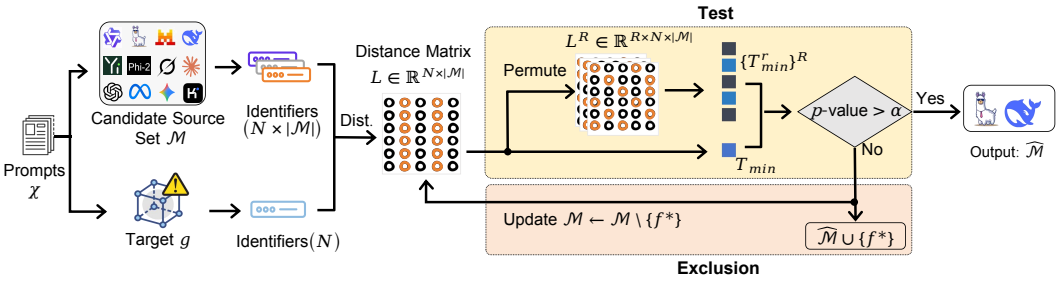


Figure 1: **Overview of MPS.** Given a target model g and a candidate source set $\mathcal{M}$, MPS implements a sequential test-and-exclusion procedure to build a provenance set $\hat{\mathcal{M}}$ satisfying $Pr(\mathcal{M}^* \subseteq \hat{\mathcal{M}}) \geq 1-\alpha$, where $\mathcal{M}^* \subseteq \mathcal{M}$ is the ground-truth source set. Orange circles in L indicate the distances between g and its sources, and f^* is the most similar candidate model.

Assumption 3.1 ensures exchangeability (not independence) over distances from independent models that does not share derivation (e.g., same ancestry), consistent with the premises in prior work (Zhu et al., 2025; Kuditipudi et al., 2025). Yet this may be violated by homogeneous negative candidates (i.e., sharing the same family or architecture); see Section G.3 for empirical diagnostics on the impact of same-family candidates. To mitigate this, we use well-documented candidate models to avoid potential shared derivation relationships among them. We provide a empirical support for exchangeability between independent models in Appendix D.

We then adopt a permutation test to compute the p-value, since it preserves all observed samples in each resample and naturally leverages exchangeability under the null. Other resampling methods, such as bootstrap, are also applicable. Specifically, in each permutation round r, we permute the candidate-wise distances independently within every prompt across the entire score matrix, and then recompute the t-statistics to obtain a new $T^{(r)}_{\min}$. Repeating for R rounds yields a reference distribution $\{T^{(r)}_{\min}\}_{r=1}^{R}$, where the p-value is computed as

$$p_{\mathcal{M}} = \frac{1 + \sum_{r=1}^{R} \mathbf{1}\left\{T^{(r)}_{\min} \leq T^{\text{obs}}_{\min}\right\}}{R+1}, \tag{9}$$

where $T^{obs}_{\min}$ is computed from the unpermuted L. The $p_{\mathcal{M}}$ quantifies the likelihood of $T^{obs}_{\min}$ under the null stage and typically determines the statistical significance of the observed result by comparing with a predetermined significance level α. Following this criterion, we reject the null stage whenever $p_{\mathcal{M}} \leq \alpha$, indicating that the observed minimal distance is statistically significant (i.e., too small to be explained by random chance). Otherwise, when $p_{\mathcal{M}} > \alpha$, we fail to reject the null stage and conclude that no remaining model can be statistically distinguished as being closer to g.

Theorem 3.1 (Validity permutation p-value)**.** *For any candidate set $\mathcal{M}$, suppose that (a) prompts $\{x_t\}_{t=1}^{N}$ are sampled independently, (b) $\mathbb{E}[|d_{i\cdot,t}|^{2+\delta}] < \infty$ for some $\delta > 0$ and every $f_i \in \mathcal{M}$, (c) $\sigma_i^2(\mathcal{M}) = \mathrm{Var}(d_{i\cdot,t}) > 0$, and (d) Assumption 3.1 holds for $\mathcal{M}$. Then the p-value $p_{\mathcal{M}}$ computed in Eq. (9) is valid for the row-wise permutation test. In particular,*

$$\limsup_{N,R\to\infty} \Pr(p_{\mathcal{M}} \leq \alpha) \leq \alpha. \tag{10}$$

Condition (a) ensures independence across samples, (b) ensures the central limit theorem applies, (c) rules out degenerate cases where two models produce identical outputs, and (d) gives the row-wise null invariance used by the permutation test. These are regular conditions applied to testing problems. The proof of Theorem 3.1 appears in Appendix B.3.

3.3 Model Provenance Set Construction

To identify all sources, MPS iteratively refines the candidate set $\mathcal{M}$ via test-and-exclusion (Algorithm 1); see Appendix C.1 for complexity analysis. The procedure maintains a sequential structure of the test, avoiding error accumulation across iterations. Specifically, at each step, it tests via p-value whether all candidates are statistically indistinguishable in similarity to g; if not, it moves the most

similar candidate to $\hat{\mathcal{M}}$, repeating until the null is no longer rejected. The final $\hat{\mathcal{M}}$ contains the source models identified by MPS. For the special case with no true source($|\mathcal{M}^*| = \emptyset$), under the expected behavior, MPS would accept the null hypothesis at first iteration and returns an empty set.

By construction, MPS offers several attractive advantages. One is that it naturally accommodates multiple candidates and multiple true sources, supporting a broader range of provenance scenarios beyond the pairwise setting. Another is that it requires no external reference models, as it directly exploits the relative fingerprint distances to detect provenance signals, without requiring carefully constructed references to provide decision boundaries. Most critically, it delivers provable coverage guarantees while producing a compact provenance set bounded at risk level α. Even in complex lineage scenarios where provenance signals are subtle, a sufficiently large query space and permutation rounds theoretically enables valid null distribution construction, allowing true provenance to be distinguished from noise. We now formally establish these theoretical guarantees as follow.

Proposition 3.1 (Power and correct exclusion)**.** *With Assumption 2.1 and the conditions of Theorem 3.1, for any nonempty $\mathcal{R} \subseteq \mathcal{M}^*$ and $\mathcal{U} = \mathcal{M} \setminus \mathcal{M}^*$, let $\mathcal{M}_\mathcal{R} = \mathcal{R} \cup \mathcal{U}$ be a active set that evolves with iterations. Then as $N, R \to \infty$,*

$$\Pr\{p_{\mathcal{M}_\mathcal{R}} \leq \alpha\} \to 1, \tag{11}$$

$$\Pr\big(arg \min_{f_i \in \mathcal{M}_\mathcal{R}} t_i(\mathcal{M}_\mathcal{R}) \in \mathcal{R}\big) \to 1 \tag{12}$$

Theorem 3.2 (Coverage guarantee)**.** *Let Assumption 2.1 holds; then with Proposition 3.1, Algorithm 1 produces a provenance set $\hat{\mathcal{M}}$ such that*

$$\Pr\left(\mathcal{M}^* \subseteq \hat{\mathcal{M}}\right) \to 1. \tag{13}$$

In particular, $\hat{\mathcal{M}}$ is an asymptotically valid level-$(1-\alpha)$ provenance set.

Theorem 3.3 (Asymptotic exactness)**.** *Under Assumption 2.1 holds, together with Theorem 3.1 and Proposition 3.1, the predicted provenance set $\hat{\mathcal{M}}$ output by Algorithm 1 satisfies*

$$\liminf_{N,R\to\infty} \Pr(\hat{\mathcal{M}} = \mathcal{M}^*) \geq 1 - \alpha. \tag{14}$$

If $\alpha = \alpha_N \to 0$ and the permutation count satisfies $R_N \alpha_N \to \infty$, then $\Pr(\hat{\mathcal{M}} = \mathcal{M}^*) \to 1$.

Remark 3.1 (Minimum detectable gap)**.** *For reliable detection, the standardized stepwise gap should exceed the sampling noise level. Assumption 2.1 expresses this as $\sqrt{N}\Delta_N \to \infty$. Hence, the minimum detectable standardized gap is of order $N^{-1/2}$, which matches the standard parametric rate. Under the sufficient condition, this corresponds to a raw mean gap larger than the relative standard error of the fingerprint distance.*

Proposition 3.1 ensures, at each iteration, the null is rejected whenever a ground-truth source remains (power), and the excluded model always statistically belongs to $\mathcal{M}^*$ (correct exclusion). Theorem 3.2 assembles it inductively to satisfy the coverage constraint in Eq. (5), and Theorem 3.3 guarantees the compactness of $\hat{\mathcal{M}}$ by controlling false inclusions. The proofs are in Appendix B.4–B.6.

4 Experiments

4.1 Setup

Models and benchmark. Since public provenance benchmarks are primarily designed for pairwise binary determination, they do not directly support our multi-candidate setting with source-set ground truth. We therefore construct instances of the form $(g, \mathcal{M}, \mathcal{M}^*)$. To ensure sufficient sample size, we collect LLMs and their fine-tuning derivation chains from Hugging Face (Wolf et al., 2019) using the `base_model` field in model cards, yielding 562 models with lineage depths up to three and parameter counts ranging from 135M to 14B. Each chain forms a test instance, where we treat the last LLM as the target model, the earlier ones are sources model ($|\mathcal{M}^*|$ = chain depth). For each chain, we sample $|\mathcal{M}| - |\mathcal{M}^*|$ unrelated models of g to build a candidate pool of size $|\mathcal{M}|$. For special case $|\mathcal{M}^*| = 0$ (no source), instances are constructed from chains of depth 1 by randomly sampling $|\mathcal{M}|$ unrelated models per chain. Note that unrelated models are collected from different families than the

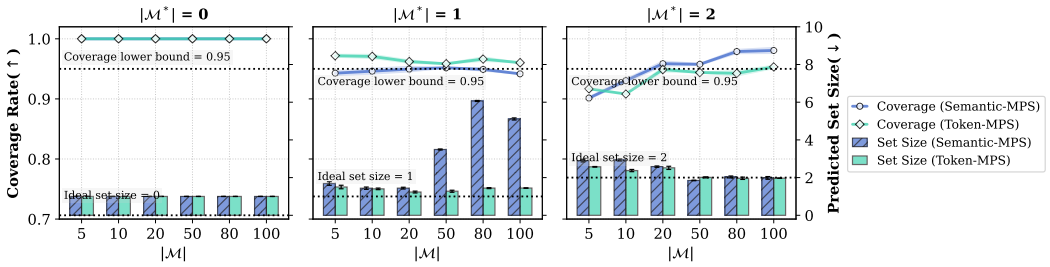


Figure 2: Provenance performance of Semantic-MPS and Token-MPS at significance level $\alpha = 0.05$. $|\mathcal{M}|$ is the candidate pool size, and $|\mathcal{M}^*|$ is the number of ground-truth sources per pool.

true sources and are mutually non-derived, avoiding relatedness that could violate exchangeability. We omit $|\mathcal{M}^*| = 3$ due to insufficient samples. Table 1 summarizes the benchmark; see Appendix E for details on derivation chain collection and test construction.

Evaluation metrics. We report two typical metrics in set prediction tasks, which also directly connect to the objective in Section 2: (1) **Coverage rate**, which measures the fraction of instances whose all source models are completely identified. By design, it is lower bounded by $1 - \alpha$. (2) **Predicted set size**, which is closer to ground-truth size $|\mathcal{M}^*|$ indicating better selection.

Framework. We implement two established indicators to measure provenance relationships: (1) next-token distance (Nikolic et al., 2025) for linguistic divergence on a curated set of approximately 100k prompts, referred to as Random100k, and (2) response semantic distance (Wu et al., 2025b) for semantic dissimilarity on general tasks. Since the latter imposes no prompt constraint, we use Random100k for both measurements. For clarity, we refer to the resulting variants as *Token-MPS* and *Semantic-MPS*, respectively.

Table 1: Benchmark statistics grouped by model lineages depth. "-" indicates omitted settings due to insufficient derivation chains for reliable evaluation.

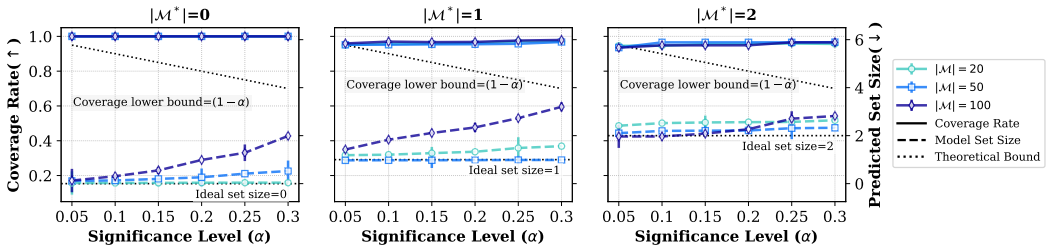
Feature	Depth=1	Depth=2	Depth=3
Base models	130	64	40
Derived models	456	132	28
Derivation chains	462	132	28
Test instances	462	132	-
Ground truths	0 / 1	2	-

Implementation details. In the MPS procedure, each model is asked to response via greedy decoding, ensuring stable behavioral patterns to suppress stochastic noise. All experiments are conducted with a significance level of $\alpha = 0.05$, permutation count $R = 1000$, and prompt count $T = 100$k, unless otherwise stated. We report the averaged values of each metric across 10 runs. More details is provided in Appendix F.

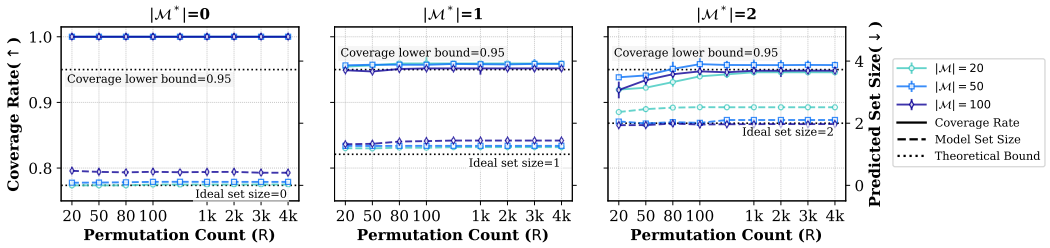
4.2 Main Results

MPS enables coverage of all true provenances with near-minimal sets. Figure 2 presents the performance of MPS across varying $|\mathcal{M}| \in [5, 100]$ at α = 0.05, with other significance levels reported in Appendix G.1. We observe that MPS consistently maintains coverage above the desired $1-\alpha$ bound across varying $|\mathcal{M}|$. For example, with 100 candidates, MPS still robustly covers all true sources at a rate of nearly 0.95. Meanwhile, the predicted set size remains tightly concentrated around the ideal value $|\mathcal{M}^*|$. Overall, MPS achieves reliable coverage guarantee while maintaining lower false inclusions across varying candidate pool scales.

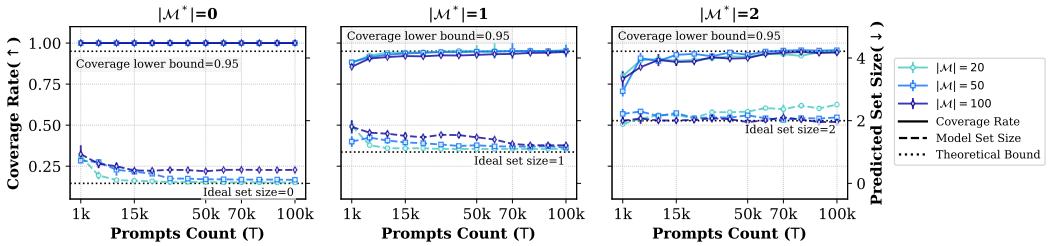
MPS benefits from larger candidate pools in deep lineage settings. As $|\mathcal{M}|$ increases, Figure 2 reveals an interesting divergence: performance decays for $|\mathcal{M}^*| = 1$ yet improves for $|\mathcal{M}^*| = 2$ in both coverage and set size. For instance, coverage for the latter clearly rises from 0.9 to above 0.95 as the pool grows, while for the former it stays around 0.95. We attribute this shift to lineage signal attenuation, where provenance signals weaken as lineage depth increases (see Appendix G.2 for detailed analysis). In larger pools, the inclusion of more unrelated models tends to yield a more precise estimate of the null distribution, allowing MPS to better distinguish subtle but real provenance signals from noise. More evidence is shown in Appendix G.3. Collectively, this scale-driven refinement highlights the feasibility of MPS for large-scale model auditing in complex lineage.



(a) Results for varying significance level α



(b) Results for varying permutation count R



(c) Results for varying prompt count T

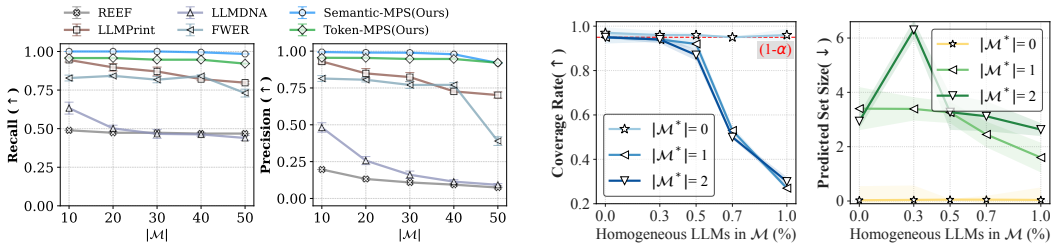
Figure 3: Sensitivity analysis of Token-MPS under varying (a) significance level α, (b) permutation count R, and (c) prompt count T. In each panel, $|\mathcal{M}^*|$ indicates the number of ground-truth sources.

MPS maintains statistical validity across significance levels. We also verify that MPS empirically satisfies its theoretical coverage guarantees with any significance level α. As shown in Figure 3(a) for Token-MPS and Figure 9 in Appendix G.4 for Semantic-MPS, the coverage consistently remains above the 1-α bound, providing empirical support for the statistical guarantee. Furthermore, the predicted set size grows with α as expected, since a larger α relaxes the theoretical floor and allows more tolerance for false inclusions, yet it stays relatively compact, e.g., always remaining below 4 for any a when $|\mathcal{M}| = 100$. Overall, MPS exhibits well-calibrated behavior consistent with theory, allowing practitioners to flexibly tune α without compromising statistical reliability.

MPS achieves stability with modest computational overhead. To assess the cost-effectiveness of MPS, we evaluate Token-MPS across two different overhead settings: permutation counts R and prompt counts T. As shown in Figure 3(b) and Figure 3(c) respectively, the coverage of MPS exceeds the $1 - \alpha$ bound even at a small tested scale, e.g., $R = 20$ or $T = 15\text{k}$ when $|\mathcal{M}^*| = 1$, while the prediction set size rapidly shrinks toward the ideal value as overhead increases. Specifically, both coverage rate and set size stabilize once $R \geq 100$ and T reaches $15\text{k} \sim 70\text{k}$ (depending on involved lineage depth). Note that such prompts are effortless to acquire as they are randomly generated simple sentences. Overall, MPS achieves reliable determination with relatively low computational cost, ensuring its practicality for large-scale provenance tasks.

5 Discussion

Bench-A is chosen for baseline comparison because it provides the reference models required by some baselines, while our benchmark serves as the labeled training data for classifiers like LLMDNA



(a) Provenance identification performance

(b) Performance under exchangeability violations

Figure 4: (a) Comparison with baselines in multi-candidate settings ($|\mathcal{M}| > 1$);higher recall means fewer missed sources and higher precision fewer false inclusions. (b)Performance of Token-MPS across increasing fractions of homogeneous negative candidates, with $\alpha = 0.05$ and pool size 20.

Comparison on provenance Identification. Existing methods focus on pairwise provenance identification, determining whether a candidate model f is a source of g. They can be extended to multiple candidates by testing each pair independently and aggregating positive decisions. Here, we thus study *how accurately these methods recover the true source set*. We compare MPS with representative baselines, including threshold-based REEF (Zhang et al., 2025b) and LLMPrint (Hu et al., 2025), classifier-based LLMDNA (Wu et al., 2025b), and test-based FWER (Nikolic et al., 2025). We report set-level recall and precision, which complement coverage rate by measuring how well $\hat{\mathcal{M}}$ matches $\mathcal{M}^*$. For fairness, we construct multi-candidate instances from the public Bench-A benchmark (Nikolic et al., 2025), containing 94 provenance/non-provenance pairs and 10 hold-out control models required by FWER. In Figure 4(a), MPS consistently outperforms the baselines across candidate pool sizes $|\mathcal{M}|$, with both precision and recall above 0.92. This demonstrates its effectiveness in recovering the true source set under multiple candidates. Additional implementation details and results, including coverage rate and predicted set size, are provided in Appendix H.

Robustness to exchangeability violations. Our theoretical coverage guarantee is established under an exchangeability assumption, which assumes that negative candidates do not exhibit systematic dependence. This assumption may be practically violated when negative candidates are homogeneous, e.g., sharing the same model family or architecture as the true sources of g. In this part, we study *the robustness of MPS to such violations* by progressively increasing the fraction β of homogeneous negatives. As shown in Figure 4(b), MPS remains robust under moderate violations: with up to 30% contaminated negatives, the empirical coverage remains close to the target level, while substantial degradation appears only under stronger violations. Moreover, increasing the candidate pool size further mitigates the impact of exchangeability violations, indicating that greater candidate diversity can improve robustness. Implementation details are provided in Appendix G.3.

6 Conclusion

In this work, we formulate model provenance as a set-valued inference problem with formal coverage guarantees and propose a principled method, *MPS*, that adaptively identifies source models through sequential significance testing. We establish the asymptotic theory of MPS, showing that it attains the desired coverage level while controlling false inclusions. Extensive experiments validate its effectiveness for model provenance identification, including challenging settings with strong distractor candidates (Section 5). To the best of our knowledge, this is the *first* work to study provenance set inference with provable statistical guarantees. We hope this preliminary effort encourages further research on reliable model provenance analysis and facilitates its practical deployment.

Limitation and further work. Findings from this study need to be interpreted in light of its limitations. One concern is the incompleteness of the ground-truth provenance annotations. Although these are supplied by model up-loaders on Hugging Face, they may be missing or mislabeled in practice, as developers do not always disclose the full lineage of their models. We partially alleviate this by focusing on widely recognized foundation models, though a truly reliable benchmark remains an open challenge requiring broader community effort. Future work could also explore more powerful test statistics and effective similarity metrics that remain sensitive across deeper derivation lineages.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the clarity and presentation of the paper. They were not used to generate experimental data or determine the reported numerical results. All AI-assisted content was reviewed, revised, and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

Model provenance estimates can affect model developers and downstream users if interpreted as definitive evidence of unauthorized reuse. MPS should be used as a screening tool for further verification: statistical evidence of provenance does not by itself establish a licensing violation. Its guarantees depend on the stated assumptions, and the model-card lineage annotations used for evaluation can be incomplete or incorrect, as discussed in Section 6. Any downstream audit should account for these limitations and independently examine the relevant model documentation and permissions.

REPRODUCIBILITY STATEMENT

Section 2 states the problem formulation and assumptions, and Section 3 presents the testing procedure and theoretical results. Complete proofs are provided in Appendix B. Appendix E describes the model collection and benchmark construction, while Appendix F specifies the distance functions and experimental environment. Experimental settings are described in Section 4, with baseline details in Appendix H.1. The code and data link is provided in Section 1.

REFERENCES

Alessandra Amendola, Manuela Braione, Vincenzo Candila, and Giuseppe Storti. A model confidence set approach to the combination of multivariate volatility forecasts. *International Journal of Forecasting*, 36(3):873–891, 2020.

Mauro Bernardi and Leopoldo Catania. The model confidence set package for r. *International Journal of Computational Economics and Econometrics*, 8(2):144–158, 2018.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Piersilvio De Bortoli, Davide Ferrari, Francesco Ravazzolo, and Luca Rossini. Model selection confidence sets for time series models with applications to electricity load data. *arXiv preprint arXiv:2602.16527*, 2026.

Davide Ferrari and Yuhong Yang. Confidence sets for model selection by f-testing. *Statistica Sinica*, pp. 1637–1658, 2015.

Peter R Hansen, Asger Lunde, and James M Nason. The model confidence set. *Econometrica*, 79(2): 453–497, 2011.

Yuepeng Hu, Zhengyuan Jiang, Mengyuan Li, Osama Ahmed, Zhicong Huang, Cheng Hong, and Neil Gong. Fingerprinting LLMs via prompt injection. *arXiv preprint arXiv:2509.25448*, 2025.

Amazon Artificial General Intelligence. The amazon nova family of models: Technical report and model card. 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning (ICML)*, pp. 17061–17084. PMLR, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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.

Linyang Li, Botian Jiang, Pengyu Wang, Ke Ren, Hang Yan, and Xipeng Qiu. Watermarking LLMs with weight quantization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3368–3378, 2023a.

Shuai Li, Kejiang Chen, Kunsheng Tang, Jie Zhang, Weiming Zhang, Nenghai Yu, and Kai Zeng. Turning your strength into watermark: Watermarking large language model via knowledge injection. *arXiv preprint arXiv:2311.09535*, 2023b.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 6499–6519, 2024.

Xiaohui Liu, Yuanyuan Li, and Jiming Jiang. Simple measures of uncertainty for model selection. *Test*, 30(3):673–692, 2021.

Anshul Nasery, Edoardo Contente, Alkin Kaz, Pramod Viswanath, and Sewoong Oh. Are robust LLM fingerprints adversarially robust? *arXiv preprint arXiv:2509.26598*, 2025.

Ivica Nikolic, Teodora Baluta, and Prateek Saxena. Model provenance testing for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NurIPS)*, 2025.

Shinwoo Park, Hyejin Park, Hyeseon Ahn, and Yo-Sub Han. Watermod: Modular token-rank partitioning for probability-balanced LLM watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, 2026. Oral presentation.

Dario Pasquini, Evgenios M Kornaropoulos, and Giuseppe Ateniese. LLMmap: Fingerprinting for large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 299–318, 2025.

Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun, and Xing Xie. Are you copying my model? protecting the copyright of large language models for eaas via backdoor watermark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7653–7668, 2023.

Zhenzhen Ren, GuoBiao Li, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. CoTSRF: Utilize chain of thought as stealthy and robust fingerprint of large language models. *arXiv preprint arXiv:2505.16785*, 2025.

Zhuoyi Shang, Jiasen Li, Pengzhen Chen, Yanwei Liu, Xiaoyan Gu, and Weiping Wang. Attesting model lineage by consisted knowledge evolution with fine-tuning trajectory. *arXiv preprint arXiv:2601.11683*, 2026.

Qin Shi, Amber Yijia Zheng, Qifan Song, and Raymond A. Yeh. Knowledge distillation detection for open-weights models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, Tianming Liu, Neil Zhenqiang Gong, Jiliang Tang, Caiming Xiong, Heng Ji, Philip S. Yu, and Jianfeng Gao. A survey on post-training of large language models. *arXiv e-prints*, 2025. doi: 10.48550/arXiv.2503.06072.

Faguang Wen, Jiming Jiang, and Yihui Luan. Model selection path and construction of model confidence set under high-dimensional variables. *Mathematics*, 12(5), 2024. ISSN 2227-7390.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

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.

H Xu, L Xiang, X Ma, B Yang, and B Hufu Li. A modality-agnositc watermarking system for pre-trained transformers via permutation equivariance. *arXiv preprint arXiv:2403.05842*, 2024.

Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhang Kong, Changting Lin, et al. Copyright protection for large language models: A survey of methods, challenges, and trends. *arXiv preprint arXiv:2508.11548*, 2025a.

Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. CTCC: A robust and stealthy fingerprinting framework for large language models via cross-turn contextual correlation backdoor. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6978–7000, 2025b.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024a.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024b.

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.

Gianpaolo Zammarchi and Yoshinori Kawasaki. Assessing the quality of forecasting quantitatively using model confidence set: from naive models to neural networks. *Quality & Quantity*, pp. 1–25, 2026.

Denghui Zhang, Zhaozhuo Xu, and Weijie Zhao. LLMs and copyright risks: Benchmarks and mitigation approaches. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pp. 44–50, 2025a.

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

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation encoding fingerprints for large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 48092–48117, 2025b.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. REMARK-LLM: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, 2024.

Chao Zheng, Davide Ferrari, and Yuhong Yang. Model selection confidence sets by likelihood ratio testing. *Statistica Sinica*, 29(2):827–851, 2019.

Sally Zhu, Ahmed M Ahmed, Rohith Kuditipudi, and Percy Liang. Independence tests for language models. In *International Conference on Machine Learning (ICML)*, pp. 79673–79698. PMLR, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Related work

This section reviews three lines of work relevant to our provenance set. First, we review existing methods for identifying model relationships, including model provenance detection and ownership verification. We then discuss recent advances in model selection, particularly model confidence sets for uncertainty quantification.

Model provenance detection. Existing methods distinguish provenance-related model pairs from unrelated ones using similarity measures, and can be broadly divided into white-box and black-box approaches. **White-box** methods exploit model parameters or internal representations, such as attention parameter similarity, invariants (Yoon et al., 2025), or hidden representations (Hu et al., 2025; Zhang et al., 2025b). **Black-box** methods instead rely on model outputs, including semantic or functional similarity (Wu et al., 2025b) and behavioral features (Kuditipudi et al., 2025; Ren et al., 2025). These methods typically use heuristic rules such as thresholding or learned classifiers and lack formal statistical guarantees. Recent work introduces hypothesis testing for more rigorous provenance detection. For example, Nikolic et al. (2025) uses a z-test to detect significant next-token similarity relative to reference models, while Zhu et al. (2025) tests hidden-activation similarity. However, these methods remain pairwise; independently testing multiple candidates can accumulate errors and does not directly provide set-level guarantees. In contrast, our work jointly reasons over the candidate set and selects a provenance subset with provable error control.

Model ownership verification. Model ownership verification establishes whether a model belongs to a particular owner using distinctive identifiers. **Watermarking** embeds signatures during training through parameter modification (Li et al., 2023b; Xu et al., 2024; Li et al., 2023a) or at inference time through latent representation (Zhang et al., 2024; Peng et al., 2023) or decoding modifications (Dathathri et al., 2024; Kuditipudi et al., 2024; Kirchenbauer et al., 2023; Liu et al., 2024; Park et al., 2026). These methods require proactive intervention and may be weakened by subsequent fine-tuning. In contrast, **fingerprinting** identifies models post hoc from internal properties (Wu et al., 2025a; Hu et al., 2025) or output behaviors (Pasquini et al., 2025; Kuditipudi et al., 2025; Xu et al., 2025b). These methods primarily verify model identity rather than recover complete derivation lineages. Although our framework is agnostic to the choice of similarity score, such identity-based scores may not always be suitable for provenance identification. For example, models from the same organization may share similar identity features without a derivation relationship, potentially confounding provenance detection.

Model confidence set. The model confidence set (MCS) (Hansen et al., 2011) quantifies model-selection uncertainty by returning a set containing the best-performing model at a prescribed confidence level. It typically combines equivalence tests with elimination rules, expanding the predicted set when data are insufficient to distinguish models (Bernardi & Catania, 2018; Zammarchi & Kawasaki, 2026). Extensions include variable selection confidence sets based on F-tests (Ferrari & Yang, 2015) or likelihood ratio tests (Zheng et al., 2019; De Bortoli et al., 2026), as well as model-level methods for portfolio selection (Amendola et al., 2020), bootstrap rank cutting (Liu et al., 2021), and model averaging (Wen et al., 2024). These methods assume that the candidate set contains at least one best-performing model, which may not hold in provenance settings where no true ancestor exists. We instead develop a set-valued provenance framework in the inverse direction, which can handle zero, single, or multiple true ancestors with provable coverage guarantees.

B Proofs

B.1 A Sufficient Condition for Separability

We formally state the sufficient condition underlying Assumption 2.1 below.

Lemma B.1 (Mean-gap sufficient condition)**.** *Suppose $\mathcal{U} \neq \emptyset$ and the following two conditions hold: (i) **Mean-gap condition**. There exists a constant $\delta > 0$ such that $\max_{f_i \in \mathcal{M}^*} \mu_i \leq \min_{f_u \in \mathcal{U}} \mu_u - \delta$. (ii) **Common-variance condition**. For every nonempty subset $\mathcal{R} \subseteq \mathcal{M}^*$, the relative variances are all equal within $\mathcal{M}_\mathcal{R} = \mathcal{R} \cup \mathcal{U}$, meaning that $\sigma_i(\mathcal{M}_\mathcal{R}) = \sigma(\mathcal{M}_\mathcal{R})$ for all $f_i \in \mathcal{M}_\mathcal{R}$, and also $\sigma(\mathcal{M}_\mathcal{R}) \leq \bar{\sigma} < \infty$. Then Assumption 2.1 holds with a constant-order separability signal $\Delta_N = \delta/(M\bar{\sigma}) = O(1)$.*

We provide the proof of Lemma B.1 below:

Proof. Fix any nonempty $\mathcal{R} \subseteq \mathcal{M}^*$ and let $\mathcal{M}_\mathcal{R} = \mathcal{R} \cup \mathcal{U}$. Let $f_\mathcal{R} \in \arg\min_{f_i \in \mathcal{R}} \mu_i$ be the source model with the smallest expected distance to g within $\mathcal{R}$. By the mean-gap condition, for every $f_u \in \mathcal{U}$,

$$\mu_u - \mu_{f_\mathcal{R}} \geq \min_{f_u \in \mathcal{U}} \mu_u - \max_{f_i \in \mathcal{M}^*} \mu_i \geq \delta.$$

Since $\mathcal{U} \neq \emptyset$ and $|\mathcal{M}_\mathcal{R}| \leq M$, the ensemble mean satisfies

$$\frac{1}{|\mathcal{M}_\mathcal{R}|} \sum_{f_j \in \mathcal{M}_\mathcal{R}} \mu_j = \mu_{f_\mathcal{R}} + \frac{1}{|\mathcal{M}_\mathcal{R}|} \sum_{f_j \in \mathcal{M}_\mathcal{R}} (\mu_j - \mu_{f_\mathcal{R}}) \geq \mu_{f_\mathcal{R}} + \frac{|\mathcal{U}|\delta}{|\mathcal{M}_\mathcal{R}|} \geq \mu_{f_\mathcal{R}} + \frac{\delta}{M},$$

where the last inequality uses $|\mathcal{U}|/|\mathcal{M}_\mathcal{R}| \geq 1/M$. Therefore, the centered distance satisfies

$$\theta_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R}) = \mu_{f_\mathcal{R}} - \frac{1}{|\mathcal{M}_\mathcal{R}|} \sum_{f_j \in \mathcal{M}_\mathcal{R}} \mu_j \leq -\frac{\delta}{M}.$$

Applying the common-variance condition $\sigma_i(\mathcal{M}_\mathcal{R}) = \sigma(\mathcal{M}_\mathcal{R}) \leq \bar{\sigma}$ yields

$$\gamma_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R}) = \frac{\theta_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R})}{\sigma(\mathcal{M}_\mathcal{R})} \leq -\frac{\delta}{M\bar{\sigma}},$$

establishing the first inequality in Assumption 2.1 with $\Delta_N = \delta/(M\bar{\sigma})$. For every $f_u \in \mathcal{U}$, since $\sigma_i(\mathcal{M}_\mathcal{R}) = \sigma(\mathcal{M}_\mathcal{R})$ for all $f_i \in \mathcal{M}_\mathcal{R}$,

$$\gamma_u(\mathcal{M}_\mathcal{R}) - \gamma_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R}) = \frac{\theta_u(\mathcal{M}_\mathcal{R}) - \theta_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R})}{\sigma(\mathcal{M}_\mathcal{R})} = \frac{\mu_u - \mu_{f_\mathcal{R}}}{\sigma(\mathcal{M}_\mathcal{R})} \geq \frac{\delta}{\bar{\sigma}} \geq \frac{\delta}{M\bar{\sigma}},$$

where the second equality follows from the cancellation of the ensemble mean in the difference $\theta_u - \theta_{f_\mathcal{R}}$, and the first inequality applies the mean-gap condition. Therefore,

$$\gamma_{f_\mathcal{R}}(\mathcal{M}_\mathcal{R}) \leq \min_{f_u \in \mathcal{U}} \gamma_u(\mathcal{M}_\mathcal{R}) - \frac{\delta}{M\bar{\sigma}},$$

establishing the second inequality in Assumption 2.1 with the same $\Delta_N = \delta/(M\bar{\sigma})$. Since both inequalities hold for the same $f_\mathcal{R} \in \mathcal{R}$ and the same constant $\Delta_N = \delta/(M\bar{\sigma}) > 0$, and since $\sqrt{N}\Delta_N \to \infty$ as Δ_N is a positive constant independent of N, Assumption 2.1 holds with a constant-order separability signal. □

B.2 Population Characterization under Mean-Gap Separability

We now state a proposition that describes $\mathcal{M}^*$ in a statistically identifiable formulation.

Proposition B.1 (Population characterization under mean-gap separability)**.** *Under the sufficient condition in Lemma B.1, the provenance set including all sources admits the following population-level characterization:*

$$\mathcal{M}^* = \left\{ f_i \in \mathcal{M} : \mu_i < \min_{f_u \in \mathcal{U}} \mu_u \right\}.$$

That is, at the population level, source models are exactly those candidates whose expected distance to the target g is strictly smaller than that of any unrelated model.

We provide the formal proof of Proposition B.1 below.

Proof. Define $\tau_\mathcal{U} = \min_{f_u \in \mathcal{U}} \mu_u$ as the minimum expected distance among all unrelated candidates. The mean-gap condition gives

$$\max_{f_i \in \mathcal{M}^*} \mu_i \leq \tau_\mathcal{U} - \delta < \tau_\mathcal{U},$$

so every $f_i \in \mathcal{M}^*$ satisfies $\mu_i \leq \max_{f_i \in \mathcal{M}^*} \mu_i < \tau_\mathcal{U}$, establishing the forward inclusion

$$\mathcal{M}^* \subseteq \{ f_i \in \mathcal{M} : \mu_i < \tau_\mathcal{U} \}.$$

For the reverse inclusion, suppose $f_i \in \mathcal{M}$ satisfies $\mu_i < \tau_\mathcal{U}$. If $f_i \notin \mathcal{M}^*$, then $f_i \in \mathcal{U}$ by definition of $\mathcal{U} := \mathcal{M} \setminus \mathcal{M}^*$, which implies $\mu_i \geq \min_{f_u \in \mathcal{U}} \mu_u = \tau_\mathcal{U}$, contradicting $\mu_i < \tau_\mathcal{U}$. Hence $f_i \in \mathcal{M}^*$, giving the reverse inclusion

$$\{ f_i \in \mathcal{M} : \mu_i < \tau_\mathcal{U} \} \subseteq \mathcal{M}^*.$$

Combining both inclusions yields $\mathcal{M}^* = \{ f_i \in \mathcal{M} : \mu_i < \tau_\mathcal{U} \}$, which is the desired equality. □

B.3 Proof of Theorem 3.1

Proof. For a given candidate set $\mathcal{M}$, let $L \in \mathbb{R}^{N\times|\mathcal{M}|}$ be its corresponding distance matrix toward the target g. With the Assumption 3.1, the distances of L are exchangeable across models for each prompt, formulated as

$$L \overset{d}{=} \Pi(L).$$

Hence, the observed statistic $T_{\min}(L)$ and its prompt-wise permuted versions are distributionally symmetric under the null. If the full prompt-wise permutation group is used, the rank of the observed statistic among its permutation orbit is uniform up to ties, so the corresponding permutation p-value is super-uniform. For Monte Carlo permutations, the plus-one correction in Eq. (9) gives the standard randomized permutation test and preserves super-uniformity over the joint randomness of the data and the sampled permutations. Therefore, when exact prompt-wise exchangeability holds,

$$\Pr(p_{\mathcal{M}} \le \alpha) \le \alpha.$$

If prompt-wise exchangeability is imposed only through an asymptotic exchangeability condition on the studentized statistics, the moment and variance conditions give

$$(t_i)_{f_i\in\mathcal{M}} = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\left(\frac{d_{i\cdot,t} - \mathbb{E}d_{i\cdot,t}}{\sigma_i(\mathcal{M})}\right)_{f_i\in\mathcal{M}} + o_p(1).$$

The multivariate CLT then gives a Gaussian limit. When this limiting null law is invariant to model-label permutations, the map $(t_i)_i \mapsto \min_i t_i$ is continuous, so the continuous mapping theorem gives the same asymptotic exchangeability for $T_{\min}$. It follows that the row-wise permutation distribution consistently approximates the null distribution, and

$$\limsup_{N,R\to\infty} \Pr(p_{\mathcal{M}} \le \alpha) \le \alpha.$$

□

B.4 Proof of Proposition 3.1

Proof. Fix a nonempty provenance set $\mathcal{R} \subseteq \mathcal{M}^*$ and let $\mathcal{M}_{\mathcal{R}} = \mathcal{R} \cup \mathcal{U}$. For each $f_i \in \mathcal{M}_{\mathcal{R}}$, the law of large numbers and variance consistency imply

$$\bar{d}_{i\cdot}(\mathcal{M}_{\mathcal{R}}) \xrightarrow{p} \theta_i(\mathcal{M}_{\mathcal{R}}), \qquad \widehat{\mathrm{Var}}(\bar{d}_{i\cdot}(\mathcal{M}_{\mathcal{R}}))^{1/2} = \frac{\sigma_i(\mathcal{M}_{\mathcal{R}})}{\sqrt{N}} + o_p(N^{-1/2}).$$

Thus, uniformly over the finite set $\mathcal{M}_{\mathcal{R}}$,

$$t_i(\mathcal{M}_{\mathcal{R}}) = \sqrt{N}\gamma_i(\mathcal{M}_{\mathcal{R}}) + o_p(\sqrt{N}).$$

The first inequality in Assumption 2.1 gives that there exists $f_{\mathcal{R}} \in \mathcal{R}$ such that $\gamma_{f_{\mathcal{R}}}(\mathcal{M}_{\mathcal{R}}) \le -\Delta_N$. Hence

$$t_{f_{\mathcal{R}}}(\mathcal{M}_{\mathcal{R}}) \le -\sqrt{N}\Delta_N + o_p(\sqrt{N}) \to -\infty.$$

The permutation null distribution is tight under the regularity conditions, so its critical values are $O_p(1)$. Since the observed $T_{\min}(\mathcal{M}_{\mathcal{R}}) \le t_{f_{\mathcal{R}}}(\mathcal{M}_{\mathcal{R}}) \to -\infty$, the permutation p-value satisfies $p_{\mathcal{M}_{\mathcal{R}}} \to 0$ in probability. Therefore $\Pr\{p_{\mathcal{M}_{\mathcal{R}}} \le \alpha\} \to 1$ for any fixed $\alpha > 0$. It remains to show correct exclusion. If $\mathcal{U} = \emptyset$, the claim is immediate. Otherwise, the second inequality in Assumption 2.1 is required and gives that the same $f_{\mathcal{R}}$ satisfies

$$\gamma_{f_{\mathcal{R}}}(\mathcal{M}_{\mathcal{R}}) \le \min_{f_u\in\mathcal{U}} \gamma_u(\mathcal{M}_{\mathcal{R}}) - \Delta_N.$$

Combining this inequality with $t_i(\mathcal{M}_{\mathcal{R}}) = \sqrt{N}\gamma_i(\mathcal{M}_{\mathcal{R}}) + o_p(\sqrt{N})$ gives

$$\Pr\left(\min_{f_i\in\mathcal{R}} t_i(\mathcal{M}_{\mathcal{R}}) < \min_{f_u\in\mathcal{U}} t_u(\mathcal{M}_{\mathcal{R}})\right) \to 1.$$

Thus, the model removed by $\arg\min_{f_i\in\mathcal{M}_{\mathcal{R}}} t_i(\mathcal{M}_{\mathcal{R}})$ belongs to $\mathcal{R}$ with probability approaching one. □

B.5 Proof of Theorem 3.2

Proof. For brevity, let $K = |\mathcal{M}^*|$. If $K = 0$, then $\mathcal{M}^* \subseteq \hat{\mathcal{M}}$ holds trivially. Consider $K > 0$. By Assumption 2.1, the reference-free setting has at least 1 unrelated candidate, so every stage with a nonempty remaining provenance set has an active set of size at least 2.

We prove by *induction* that the first K exclusions are ground-truth source models with probability approaching one. We first introduce some notation for convenience. Since each iteration will exclude one potential source model, we denote the iteration index as k, where $k = 0$ indicates the initial state (no iteration performed), and $k \geq 1$ denotes the number of iterations completed. Let $\mathcal{R}_k$ be the set of ground-truth source models remaining after k exclusions, and thus $\mathcal{R}_0 = \mathcal{M}^*$.

Proposition 3.1 gives $\Pr(p_{\mathcal{M}_{\mathcal{R}_0}} \leq \alpha) \to 1$ and shows that the excluded model lies in $\mathcal{R}_0$ with probability approaching one. Now suppose that the first $k < K$ exclusions are ground-truth source models, then $\mathcal{R}_k \neq \emptyset$, and the current active set is $\mathcal{M}_{\mathcal{R}_k} = \mathcal{R}_k \cup \mathcal{U}$. Applying Proposition 3.1 again, the algorithm rejects with probability approaching one and removes a model in $\mathcal{R}_k$ with probability approaching one. Since K is finite, a union bound over $k = 0, \ldots, K-1$ gives

$$\Pr\left(\text{the first } K \text{ exclusions all belong to } \mathcal{M}^*\right) \to 1.$$

On this event all ground-truth source models are included in $\hat{\mathcal{M}}$, so

$$\Pr(\mathcal{M}^* \subseteq \hat{\mathcal{M}}) \to 1.$$

□

B.6 Proof of Theorem 3.3

Proof. Let E_N be the event that all true provenance models are selected before any unrelated model is selected. By the proof of Theorem 3.2, $\Pr(E_N) \to 1$. On E_N, after all true provenance models have been selected, the remaining active candidate set is $\mathcal{U}$. If $|\mathcal{U}| \leq 1$, Algorithm 1 stops and no unrelated model is added. If $|\mathcal{U}| > 1$, the remaining stage is a null stage because all active candidates are unrelated to g. By Theorem 3.1,

$$\limsup_{N,R\to\infty} \Pr(p_{\mathcal{U}} \leq \alpha) \leq \alpha.$$

The algorithm can add an unrelated model only if this final null stage is rejected. Therefore,

$$\limsup_{N,R\to\infty} \Pr(\hat{\mathcal{M}} \neq \mathcal{M}^*) \leq \limsup_{N,R\to\infty} \Pr(E_N^c) + \limsup_{N,R\to\infty} \Pr(p_{\mathcal{U}} \leq \alpha) \leq \alpha.$$

Equivalently,

$$\liminf_{N,R\to\infty} \Pr(\hat{\mathcal{M}} = \mathcal{M}^*) \geq 1 - \alpha.$$

If $\alpha = \alpha_N \to 0$ and $R_N \alpha_N \to \infty$, the p-value is fine enough for the final null test, and Theorem 3.1 gives $\Pr(p_{\mathcal{U}} \leq \alpha_N) \to 0$. Together with $\Pr(E_N) \to 1$, this yields $\Pr(\hat{\mathcal{M}} = \mathcal{M}^*) \to 1$. □

C MPS Pseudocode

Algorithm 1 summarizes the proposed MPS procedure. Starting from the full candidate set, MPS computes the sample-wise distances between each candidate model and the target model. At each iteration, it performs a joint test over the current candidate set using the minimum test statistic, with the corresponding p-value estimated by permutation. If the joint null is rejected, the candidate with the smallest test statistic is added to the provenance set and removed from further consideration. This test-and-exclusion process is repeated until the joint test is no longer significant, after which all selected candidates are returned as $\hat{\mathcal{M}}$. By testing the candidate set jointly rather than independently examining each pair, MPS supports set-level coverage control during adaptive provenance inference.

Algorithm 1 Model provenance set (MPS)

Input: Candidate set $\mathcal{M}_0 = \{f_1, \ldots, f_M\}$, target model g, prompts $\mathcal{X} = \{x_1, \ldots, x_N\}$, significance level α, permutation count R, and distance function $\mathcal{L}(\cdot, \cdot)$.

Output: Predicted provenance set $\hat{\mathcal{M}}$.

Compute $L \in \mathbb{R}^{N \times |\mathcal{M}|}$ by Eq. (2)
$\mathcal{M} \leftarrow \mathcal{M}_0$, $\hat{\mathcal{M}} \leftarrow \emptyset$
while $|\mathcal{M}| > 1$ **do**
  Compute t_i for all $f_i \in \mathcal{M}$ using Eq. (6)
  $T_{\min}^{\text{obs}} \leftarrow \min_{f_i \in \mathcal{M}} t_i$
  for $r = 1$ to R **do**
    Permute $\{L_{i,t}\}_{i=1}^{|\mathcal{M}|}$ for each x_t
    Compute $\{t_i^{(r)}\}_{i=1}^{|\mathcal{M}|}$
    $T_{\min}^{r} \leftarrow \min_{f_i \in \mathcal{M}} t_i^{(r)}$
  end for
  Compute $p_{\mathcal{M}}$ using Eq. (9)
  if $p_{\mathcal{M}} > \alpha$ **then**
    break
  end if
  $f^* \leftarrow \arg\min_{f_i \in \mathcal{M}} t_i$
  $\mathcal{M} \leftarrow \mathcal{M} \setminus \{f^*\}$, $\hat{\mathcal{M}} \leftarrow \hat{\mathcal{M}} \cup \{f^*\}$
end while
return $\hat{\mathcal{M}}$

C.1 Computational Complexity of MPS

We briefly analyze the computational complexity of MPS below. Let N be the number of prompts, M the number of candidate models, and C_{dist} denote the cost of computing model outputs for a single prompt. The procedure consists of two main stages: (i) the first is to **compute the distance matrix.** According to Algorithm 1, MPS first computes the distance matrix L, which requires $O(NMC_{\text{dist}})$. (ii) Next, the **test-and-elimination loop** is executed. In each iteration of the test-and-elimination loop, computing the test statistic takes $O(NM)$, and each permutation requires $O(NM)$ to reshuffle and recompute the statistic. With R permutations, this results in $O(RNM)$ per iteration. Since there are at most M iterations, the overall cost of this stage is $O(RNM^2)$.

Therefore, the total complexity of MPS is $O(NMC_{\text{dist}} + RNM^2)$.

D Empirical Support for Exchangeability

Exchangeability is a standard assumption in prior work (Kuditipudi et al., 2025; Zhu et al., 2025). In our setting, it requires that the distances of unrelated candidates remain invariant in distribution under permutations of candidate identities. We find empirical support for this assumption.

We conduct a permutation analysis ($R = 1{,}000$, $N = 100$k) on 100 instances with $|\mathcal{M}| = 20$ and $|\mathcal{M}^*| = 0$, where no candidate has a lineage relationship with g or with any other candidate, thereby approximating the exchangeable setting. Under exchangeability, the observed $T_{\min}$ should not be systematically extreme relative to its permutation null distribution. As shown in Table 2, the null distribution is nearly symmetric (skewness $= 0.056$, mean $= 0$, std $= 1$), and the observed statistic lies at the 51.1st percentile, providing empirical support for the plausibility of the assumption.

E Models and Benchmark

Model collection. We collect model derivative chains from the Hugging Face (HF) platform by HF APIs (Wolf et al., 2019). We first crawle their To ensure coverage of widely used variants, we scraped the model ID cards of the top 12,287 models by download count, and then filtered them based on fields in their metadata. Following the filtering rules: (i) we first consider models whose size meets

Table 2: Empirical evidence for exchangeability from the permutation null distribution.

Metric	Value	Interpretation
Mean / Std	-0.001 / 0.961	Permutation distribution centered at zero, unit variance
Skewness	0.056	Permutation distribution is symmetric
Observed percentile	51.1%	Observed $T_{\min}$ near the distribution center

Table 3: Top 20 source LLMs from the benchmark

Hugging Face Model	Main Domain	#Parameters (dtype)	License	#Derived Models
Qwen/Qwen2.5-1.5B	General LLM (Base)	1,543,714,304 (BF16)	Apache-2.0	47
Qwen/Qwen2.5-0.5B-Instruct	General LLM / Instruct	494,032,768 (BF16)	Apache-2.0	32
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	Reasoning / Math	1,777,088,000 (BF16)	MIT	28
Qwen/Qwen3-0.6B	General LLM (Base) / reasoning	751,632,384 (BF16)	Apache-2.0	24
meta-llama/Llama-3.1-8B-Instruct	General LLM / Instruct	8,030,261,248 (BF16)	Llama3.1	22
meta-llama/Llama-3.2-1B-Instruct	General LLM / Instruct	1,235,814,400 (BF16)	Llama3.2	22
google/gemma-2-2b	General LLM (Base)	2,614,341,888 (F32)	Gemma	21
Qwen/Qwen3-1.7B	General LLM (Base) / reasoning	2,031,739,904 (BF16)	Apache-2.0	18
Qwen/Qwen2.5-7B	General LLM (Base)	7,615,616,512 (BF16)	Apache-2.0	17
google/gemma-3-1b-it	General LLM / Instruct	999,885,952 (BF16)	Gemma	16
google/gemma-3-270m-it	General LLM / Instruct	268,098,176 (BF16)	Gemma	14
google/gemma-2b	General LLM (Base)	2,506,172,416 (BF16)	Gemma	14
Qwen/Qwen3-14B	General LLM (Base)	14768307200 (BF16)	Apache-2.0	14
google/gemma-2-2b-it	General LLM / Instruct	2,614,341,888 (BF16)	Gemma	12
openai-community/gpt2	General LLM (Base)	137,022,720 (BF32)	MIT	12
TinyLlama/TinyLlama-1.1B-Chat-v1.0	General LLM / Chat	1,100,048,384 (BF16)	Apache-2.0	11
microsoft/Phi-3-mini-4k-instruct	General LLM / Reasoning / Instruct	3,821,079,552 (BF16)	MIT	10
deepseek-ai/deepseek-coder-1.3b-base	Code LLM	1,346,471,936 (BF16)	Deepseek	10
microsoft/phi-2	General LLM / Reasoning	2,779,683,840 (F16)	MIT	9
HuggingFaceTB/SmolLM2-135M	General LLM (Small / Efficient)	134,515,008 (BF16)	Apache-2.0	8
...	...			

our hardware constraints as inferred from the field "`safetensors`" filed and that support *Transformer* and *PyTorch* according to the field "`library_name`". (ii) Then, we construct the model derivation chains according to the field "`base_model:finetune:<base_model_name>`", which indicate the direct source for current models. (iii) Finally, we exclude LLMs without this explicit indication and remove models that do not support English as indicated in their model cards. To this end, we collect 562 LLMs spanning three derivation depths and ranging from 135M to 14B, comprising 462, 183, and 28 derivation chains of length 1, 2, and 3, respectively. Note that a single long chain is decomposed into multiple overlapping segments of varying depths. Figure 5 shows examples of derivation relationships across three rounds of fine-tuning. Moreover, we list the top-20 ancestor models with the most derived models in Table 3.

Test instance construction. We construct test instances $(\mathcal{M}, g)$ with varying candidate set sizes. To control the number of true provenance models in $\mathcal{M}$, we simply extend the established derivation chains to construct candidate pools with desired counts (i.e.,$|\mathcal{M}^*|$). For example, to construct a test instance with $|\mathcal{M}| = 10$ including two ground-truth source ($|\mathcal{M}^*|$=2), we consider a derivation chain with a lineage depth of 2: $\text{grandparent model} \rightarrow \text{parent model} \rightarrow g$, and we then augment the set with 8 additional models sampled from the remaining source models that are unrelated to g.

F Implementation Details of MPS Procedure

F.1 Distance Functions

We adopt two simple yet effective metrics to measure pairwise model relationships. Both metrics require only query access to LLMs and are computationally efficient, making them suitable for large-scale provenance verification.

Next-token distance. The first metric is next-token similarity (Nikolic et al., 2025), which evaluates whether two LLMs produce identical next tokens given the same prompt. The authors reveal that fine-tuned derivative models are more likely to generate similar next-token outputs than unrelated models. To incorporate this metric into MPS, we convert it into a distance measure by defining

$$\mathcal{L}(f, g)_x = 1 - \mathbf{1}(f(x), g(x)),$$

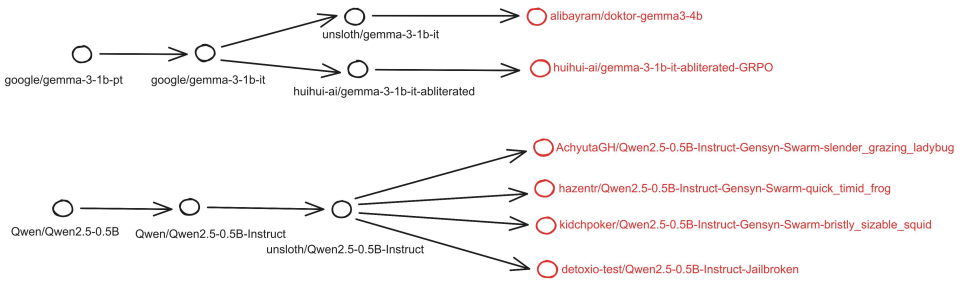


Figure 5: Examples of derivation chains with $|\mathcal{M}^*| = 3$. Red nodes means the final child LLM versions after multiple fine-tuning.

where $f(x)$ and $g(x)$ denote the predicted next tokens of LLM-f and LLM-g on prompt x, respectively.

Response semantic distance. The second metric is response embedding similarity (Wu et al., 2025b), which compares the semantic representations of model outputs generated from the same prompt. Similarly, we adapt this metric to MPS by defining the corresponding distance as

$$\mathcal{L}(f, g)_x = 1 - \cos(f(x), g(x)) \,,$$

where $f(x)$ and $g(x)$ are the embeddings of the generated outputs from LLM-f and LLM-g.

Usage in MPS procedure. For the next-token distance, given a prompt, we query each LLM to generate the next token taking the prompt as prefix input and compute the token-level discrepancy score accordingly. For the response semantic distance, following Wu et al. (2025b), we encode the generated one token of each prompt using Qwen3-Embedding-8B and use the first token embedding ([CLS]) from the final layer as the output representation.

F.2 Experimental environment

We run our model provenance testers on a Linux machine with 64-bit Ubuntu 22.04.5 LTS, 256GB RAM and 2× 16-core AMD EPYC 9124 Processor @1.50 GHz and 8× NVIDIA 4090 GPUs with 20GB RAM. All experiments are implemented using Python 3.10, CUDA 11.8, PyTorch 2.6 and Transformers 4.57.1 library.

G Additional experimental results

G.1 More results of MPS on specific significance level α

This subsection supplements a broader results of the MPS procedure under $\alpha \in \{0.1, 0.15\}$ across different score functions and candidate sizes $|\mathcal{M}| \in \{5, 10, 20, 50, 80, 100\}$. As shown in the Figure 6, the results indicate that the MPS consistently achieves a coverage rate above the predefined lower bound. Notably, across different distance scoring functions, the method maintains a compact predicted set size that remains close to the ground-truth size and is substantially smaller than the full candidate set. For instance, in the case of Semantic-MPS under $\alpha = 0.15$ with $|\mathcal{M}^*|=1$ and a candidate size $|\mathcal{M}| = 100$, the returned set size is consistently below 10, which is far less than the full candidate pool. In summary, **MPS consistently preserves prescribed coverage guarantees while returning a compact subset of candidate models**, suggesting its practical utility for large-scale provenance auditing scenarios in enabling manual or secondary verification to be concentrated on a small number of potential provenance models.

G.2 Analysis of similarity distributions

Comparing distance functions in Figures 2 and Figure 6, we observe a performance crossover: while Token-MPS is competitive at shallow lineages ($|\mathcal{M}^*|$ = 0, 1), Semantic-MPS demonstrates superior robustness as the lineage depth reaches $|\mathcal{M}^*| = 2$. **To investigate the underlying cause of this**

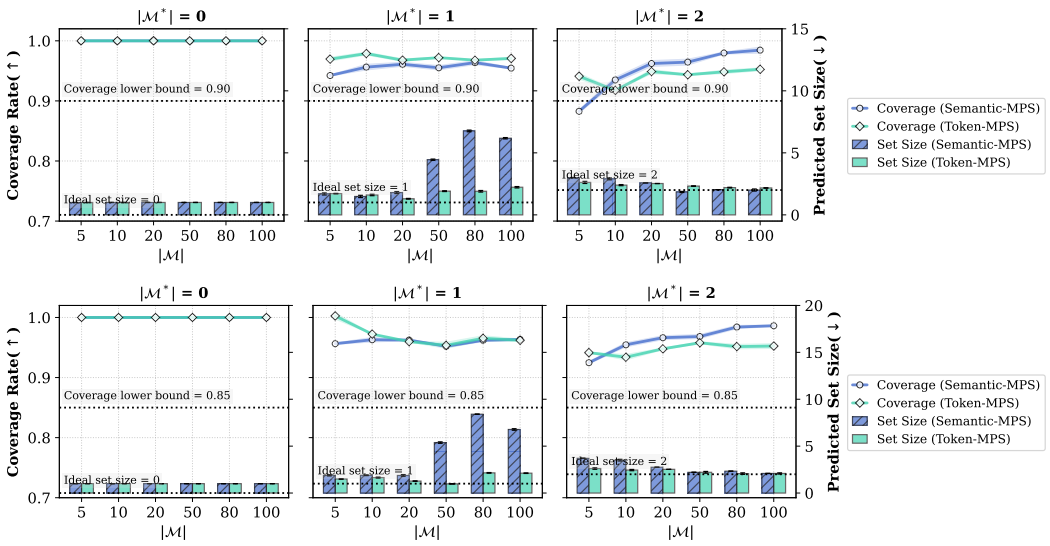


Figure 6: Averaged coverage and set size of MPS procedure with different score functions with $\alpha = 0.1$ (Top) and $\alpha = 0.15$ (Bottom). $|\mathcal{M}|$ denotes the candidate set size; $|\mathcal{M}^*|$ is the number of true ancestor models.

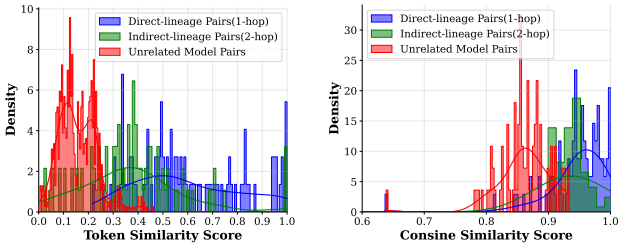


Figure 8: Similarity score distributions for unrelated pairs (red) and lineage-related pairs at different derivation distances. *1-hop* refers to parent–child models (blue), and *2-hop* indicates grandparent–grandchild (green).

performance shift, we analyze the similarity distributions across varying lineage depths. We sample 100 unrelated and lineage-related model pairs from $|\mathcal{M}^*| = 2$ derivation chains (e.g., $A \to B \to C$, where B and A represent one-hop and two-hop ancestors of C).

Figure 8 illustrates the comparative similarity distributions for Token and Semantic metrics, yielding several key insights: (i) **Similarity decays monotonically with derivation depth**. In particular, 1-hop pairs are most similar; 2-hop pairs show lower similarity for both metrics. This trend holds for both token-level and cosine semantic similarity, confirming that repeated fine-tuning progressively erodes the original feature correspondence. (ii) **Lineage-related pairs remain separable from unrelated ones.** Even at 2-hop, overlap with unrelated pairs is limited, indicating persistent provenance signals. (iii) **Token similarity excels at shallow derivations while cosine similarity handles deeper ones.** For 1-hop direct pairs, token scores show a wider gap from unrelated baselines, capturing close ties effectively. For 2-hop indirect pairs, token features drift with successive fine-tuning and overlap more with unrelated ones, whereas cosine similarity retains clearer separation, as semantic signals are better preserved. Overall, the observed performance shift suggests that token features are more susceptible to fine-tuning drift than semantic features. This motivates the development of robust scoring functions that can better capture underlying lineage continuity in complex provenance settings.

Table 4: Averaged Coverage Rate/Predicted Set Size of Token-MPS at α=0.05 and pool size=20. We bold CRs within ± 0.05 of 1.0. As all variances < 0.01, omitted for brevity.

$\lvert\mathcal{M}^*\rvert$	$\beta\% = 0.00$	$\beta\% = 0.30$	$\beta\% = 0.50$	$\beta\% = 0.70$	$\beta\% = 1.00$
0	**0.97**/ 0.03	**0.96**/ 0.04	**0.96**/ 0.05	**0.95**/ 0.05	**0.96**/ 0.04
1	**0.95**/ 3.40	**0.94**/ 3.39	**0.92**/ 3.26	0.53/ 2.45	0.27/ 1.6
2	**0.95**/ 2.94	**0.94**/ 6.30	0.87/3.27	0.50/ 3.12	0.30/ 2.63

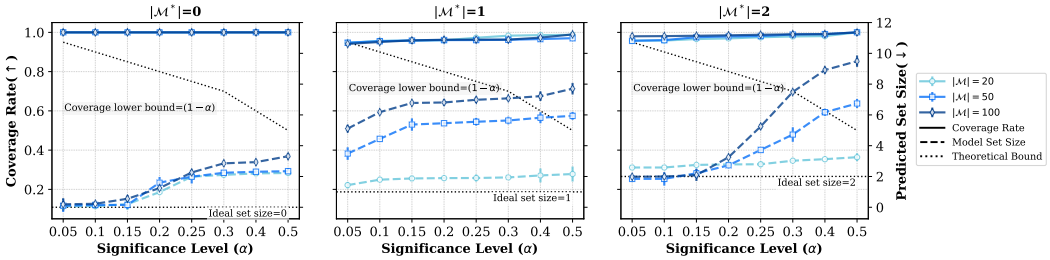


Figure 9: Average coverage rate and set size of Semantic-MPS with varying significance level α. $|\mathcal{M}|$ and $|\mathcal{M}^*|$ denote the number of candidate models and ground-true sources respectively.

G.3 Impact of same-family candidates

Here, we design a controlled experiment to empirically assess the robustness of MPS under more challenging candidate pool where the exchangeability in Assumption 3.1 may be violated by candidates sharing the same model family as the target. Note that models in the same derivation chain are treated as the same family and inherit the family label of the root ancestor. In particular, we evaluate Token-MPS on candidate pools of size 20, with systematically varying proportions ($\beta\%$) of same-family negative candidates from 0% to 100%. To ensure sufficient negative models, we consider derivation chains where the root ancestor has at least 20 same-family models (including derived ones) and construct 282 chains for $|\mathcal{M}^*| = 1/0$ and 99 chains for $|\mathcal{M}^*| = 2$, across 4 families: `Qwen`, `Google`, `Meta-Llama`, and `Deepseek-ai`. Other settings are the same as in Section 4. We report average Coverage Rate (CR) and Model Set Size (MSS) over 3 runs.

From Table 4, we observe that (i) **MPS remains robust to moderate violations of exchangeability caused by model-family overlaps.** When the portion of same-family negatives $\beta \leq 0.05$, CR stays near $1 - \alpha$, and MSS remains around $|\mathcal{M}^*|$. This shows that MPS tolerates mild exchangeability violations with mild performance degradation. (ii) **MPS avoids false inclusions under severe interference.** When($\beta \geq 0.7$), coverage drops as expected since true signals are more likely obscured by similar distractors and MPS terminates early, yet redundancy stays low, showing it is conservative. This reveals that under strong violations, MPS tends to sacrifice weak signals to avoid introducing an excess of false positives. (iii) **MPS can be strengthened by adding unrelated models.** While performance degrades with deeper derivations (where provenance signals decay), including more diverse, unrelated candidates substantially improves coverage. For example, with $|\mathcal{M}^*| = 2$, coverage jumps from 0.27($\beta = 1.0$) to 0.95($\beta = 0.0$). This illustrates that the vulnerability of MPS is not inherent but can be mitigated by enriching the candidate set with heterogeneous negatives.

G.4 Effect of significance level on Semantic-MPS

Figure 9 shows the results of Semantic-MPS under different significance levels. It can be observed that as α increases, its coverage rate consistently remains above the $1 - \alpha$ boundary. Although the increase in α leads to an expected expansion of the prediction set, the sizes remain within acceptable levels. For example, even when $|\mathcal{M}| = 100$, the set size does not exceed 10. Overall, the behaviour of MPS is statistically valid and aligns strictly with the theoretical implications of α.

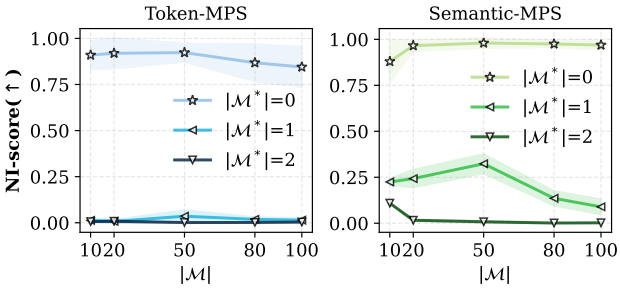


Figure 10: $p_{\mathcal{M}_i}$-value from MPS framework across different source-set size. Higher values indicate that the target model g is more likely not to infringe.

G.5 Distribution of p-value

This part analysis the first-step p-value $p_{\mathcal{M}_0}$ produced by MPS. In model IP disputes, assessing whether there is sufficient evidence of provenance requires rigorous statistical support (Xu et al., 2025a). MPS naturally provides such evidence through $p_{\mathcal{M}_0}$: when $H_{0,\mathcal{M}_0}$ cannot be rejected, the observed data provide insufficient statistical evidence that any candidate in $\mathcal{M}_0$ is related to g. Accordingly, $p_{\mathcal{M}_0}$ provides a quantitative indicator for non-infringement auditing, with larger values indicating greater consistency with the non-provenance null. Figure 10 reports the average $p_{\mathcal{M}_0}$ over 100 $(\mathcal{M}, g)$ instances at each lineage depth in our benchmark. We observe that $p_{\mathcal{M}_0}$ remains consistently above 0.85 in non-infringing scenarios ($|\mathcal{M}^*| = 0$) while dropping to near zero for infringing cases ($|\mathcal{M}^*| > 0$). This clear separation demonstrates the utility of $p_{\mathcal{M}_0}$ as a statistically grounded indicator for automated provenance auditing.

H Additional results for discussion

H.1 Baselines

We compare MPS with the following provenance methods including two white-box approaches, i.e., REEF and LLMPrint and two black-box techniques, i.e., LLMDNA and FWER.

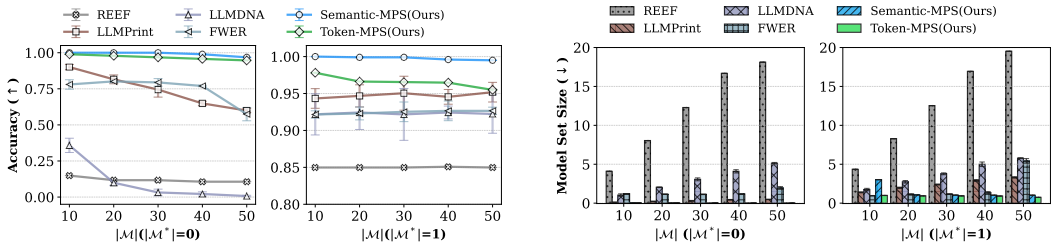
REEF (Zhang et al., 2025b) extracts intermediate hidden representations as model fingerprints and computes pairwise similarity using Centered Kernel Alignment (CKA). Following the default setting, we use the TruthfulQA dataset to extract the embeddings at layer 18 and then calculate the linear CKA score between two LLMs. As REEF does not perform provenance classification based on the CKA score, we simply adopt a threshold-based strategy based on empirical observations on the CKA scores over 281 unrelated and directly related pairs from our benchmark (excluding the overlapping pairs from Bench-A), where pairs with a score greater than 0.65 are classified as provenance.

LLMPrint (Hu et al., 2025) optimizes the fingerprint triggers by the Greedy Coordinate Gradient (GCG) algorithm (Zou et al., 2023) that generates desired adversarial suffixes in a prompt based on gradient information, and then computes a token preferences similarity by analyzing the logits of the first generated tokens between two LLMs. Following LLMPrint, we construct 300 fingerprint prompts per base LLM on the given token pairs, and use the fixed prefix "Randomly output a word from your vocabulary" followed by a suffix s_{init} with 20 placeholders ("x") for optimization. We run the GCG algorithm for 1,000 iterations to optimize a fingerprint prompt. Subsequently, provenance relationships are determined using the gray-box thresholding strategy proposed in LLMPrint with $z = 1.64$.

LLMDNA (Wu et al., 2025b) extracts semantic representations for each LLM response and then trains a support vector machine (SVM) with an RBF kernel to classify the model pairs. We join the generated tokens per LLM into a single sequence by the whitespace and then use the Qwen3-Embedding-8B to encode them. During the training, we use $|\mathcal{M}^*| = 0$ and 1 to construct 281 negative and positive training pairs. Given a pairwise embedding $(\mathbf{e}_f, \mathbf{e}_g)$, we concatenate them as the input representation $\mathbf{z}$, i.e., $\mathbf{z} = \text{concat}(\mathbf{e}_f, \mathbf{e}_g, |\mathbf{e}_f - \mathbf{e}_g|)$. To determine the hyper-parameters C and γ for the SVM, we perform a grid search over $C \in [1, 10, 50, 100]$ and $\gamma \in [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$. Based on

Table 5: Average performance for pairwise model provenance detection over three runs(mean ± std). The best results are bolded.

Method	Precision↑	Recall↑	Accuracy↑
REEF	0.89 ± 0.11	0.83 ± 0.08	0.85 ± 0.03
LLMPrint	0.92 ± 0.03	**0.97 ± 0.02**	0.95 ± 0.02
LLMDNA	0.91 ± 0.18	0.95 ± 0.10	0.93 ± 0.11
FWER	0.97 ± 0.01	0.89 ± 0.01	0.93 ± 0.04
Token-MPS	0.99 ± 0.00	0.92 ± 0.00	0.95 ± 0.00
Semantic-MPS	**0.99 ± 0.01**	0.96 ± 0.00	**0.97 ± 0.00**



(a) Averaged coverage rate

(b) Average predicted model set size

Figure 11: Provenance attribution performance across different methods under multi-candidate settings. $|\mathcal{M}|$ denotes the candidate pool size; $|\mathcal{M}^*|$ is the number of ground-truth source models.

this search, we select $C = 10$ and $\gamma = 10^{-3}$, while all other hyper-parameters are set to their default values.

FWER (Nikolic et al., 2025) is an hypothesis-test based method, which employs Z-tests under family-wise error rate (FWER) control to determine whether a suspect parent model exhibits statistically significant token-level similarity to a target model, relative to a set of control models that are known a priori to be unrelated to the target. It controls the false positive rate of the provenance determination between a pair models. Unlike joint error control in MPS, FWER runs multiple hypothesis tests independently on the (suspect parent, target) and each (control, target) model pair and alleviates error rate accumulation via the Holm–Bonferroni procedure. For a set of parent candidates, FWER outputs the candidate most similar to the target model (if it passes the significance test) or returns null. We extend this by iteratively evaluating each candidate to construct a prediction set, maintaining a significance level of $\alpha = 0.05$ and utilizing the same prompt set as MPS.

For each multi-candidate test instance $(\mathcal{M}, g)$, we apply each baseline to iteratively determine the relationship of each candidate-target pair (m, g) with $m \in \mathcal{M}$, and aggregate all candidates classified as provenance into the provenance set.

H.2 Additional results of model source attribution

Table 5 shows averaged accuracy of pairwise detection on Bench-A. Overall, MPS-based methods perform well in black-box settings, remaining competitive with white-box approaches that require internal access and carefully engineered features. Specifically, Semantic-MPS achieves the best overall performance, while Token-MPS also demonstrates strong and stable results, with all metrics exceeding 92%. Among the baselines, FWER is the closest statistically grounded comparator. However, it performs pairwise hypothesis tests by sequentially comparing the similarity between the target and a suspect source (parent) model against that of control models. This makes it difficult to distinguish weak provenance signals from control models, resulting in reduced recall (e.g., 89%). By contrast, MPS jointly analyzes relative deviations among all candidates, allowing it to better distinguish true provenance models under marginal distance differences.

Figure 11 presents averaged coverage rate and set size of different methods under multi-candidates scenario. Compared to the empirical baselines, MPS consistently outperforms across all scales,

maintaining higher accuracy (over 0.97 for Semantic-MPS and 0.95 for Token-MPS) while keeping smaller predicted set sizes near ground-truth levels. Moreover, all baselines suffer obvious accuracy degradation in the no-parent setting ($|\mathcal{M}^*| = 0$). This occurs because larger candidate sets inflate false-positive errors and produce incorrect non-empty outputs, as evidenced by the expansion of predicted set sizes.