ONE CHARACTER AT A TIME: FRUSTRATINGLY SIMPLE AND COST-EFFICIENT BLACK-BOX MODEL EQUALITY TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Users of third-party large language model APIs may have little direct evidence that the served model matches the advertised one. Black-box model equality tests offer a practical auditing mechanism, but existing methods can require many queries, long generations, or a costly query optimization stage. We study whether reliable auditing is possible with minimal resources and find a frustratingly simple yet effective equality-testing strategy that requests only one character per model response. Our method asks both the candidate and a trusted reference model to output one symbol from a fixed alphabet. Repeated responses yield empirical categorical distributions, which we compare using total variation (TV) distance and a paired permutation test. We evaluate our method using a cost-versus-statistical-power protocol, with same-model and different-model comparisons to assess Type I and Type II errors. Comprehensive empirical evaluation shows that our method achieves strong performance at substantially lower cost than existing state-of-the-art methods. We also develop a theoretical analysis of when single-symbol responses suffice for model equality testing.

1 INTRODUCTION

Large language models (LLMs) (Grattafiori et al., 2024; Yang et al., 2025a; Gemma Team, 2025) have demonstrated impressive capabilities across a wide range of real-world applications, such as autonomous agents (Yao et al., 2023; Shinn et al., 2023), medical AI (Singhal et al., 2023; Kung et al., 2023), and code generation (Chen et al., 2021; Li et al., 2022). Due to the high cost of deploying frontier models locally, many end users and researchers access them through hosted inference APIs and third-party platforms (Chen et al., 2024; Liang et al., 2023; Chiang et al., 2024). In these settings, the served model is hidden behind a black-box endpoint, and the model name reported by the provider may be the only evidence of its identity (Cai et al., 2025b; Gao et al., 2025).

This lack of transparency creates a concrete integrity risk. A provider may advertise a premium model while serving a smaller model, a quantized variant, a fine-tuned derivative, or a silently updated checkpoint (Cai et al., 2025b; Sun et al., 2025a; Zhang et al., 2026c). Such model substitution can reduce service quality, invalidate experimental comparisons, and cause users to pay for a model they do not receive (Cai et al., 2025b; Velasco et al., 2026b; Cao et al., 2026). Opaque inference pipelines can also conceal token manipulation and other billing behaviors that are difficult to verify from an invoice alone (Sun et al., 2026; Velasco et al., 2026a; Wang et al., 2025). These concerns have motivated a growing literature on auditing the identity of remotely served LLMs (Shao et al., 2025; Gao et al., 2025; Pasquini et al., 2025).

Black-box model equality testing provides a direct formulation of this auditing problem (Gao et al., 2025). Given a candidate API and a trusted reference model, the auditor queries both endpoints and tests whether their response distributions are equal (Gao et al., 2025; Zhu et al., 2026). Existing work has shown that sampled outputs can reveal model identity, provenance, or change without exposing model weights (Pasquini et al., 2025; Sun et al., 2025b; Nikolic et al., 2025). For example, MET (Gao et al., 2025) conducts a kernel two-sample test on responses from the target LLM and the reference model. RUT (Zhu et al., 2026) uses a rank-based statistic, and B3IT (Chauvin et al., 2026) searches for border inputs whose output distributions are especially sensitive to model changes. These

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

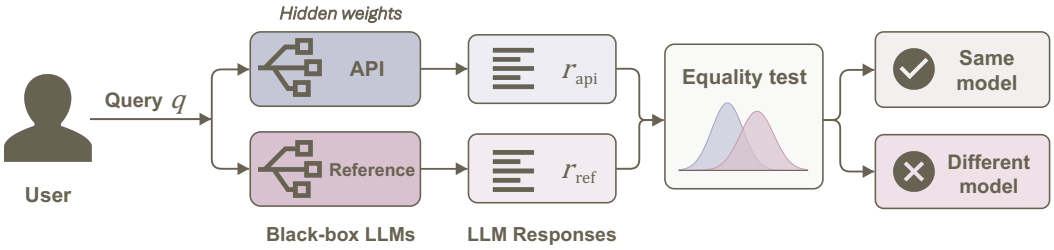


Figure 1: **Auditing LLM API identity**. A user queries the API and a reference model, then tests whether their response distributions match.

results establish feasibility, but they leave an important practical bottleneck unresolved. In order to achieve satisfactory performance, existing tests often require many queries, multiple completions, long generations, or a separate input-discovery stage, resulting in a high monetary cost per audit (Gao et al., 2025; Chauvin et al., 2026; Cai et al., 2025b). This motivates our central research question:

How can we reliably verify the identity of a black-box LLM API at the lowest financial cost?

Cost efficiency in model identity auditing matters for two reasons. First, model auditing is particularly useful when it can be repeated after model updates, routing changes, or suspected service degradation (Gao et al., 2025; Chauvin et al., 2026; Zhang et al., 2026c). A costly test is less practical for continuous monitoring across many API endpoints. Second, model auditing should also be accessible to individual users with limited budgets, enabling them to independently check the services they rely on (Shen et al., 2021; Lam et al., 2022). Affordable model verification can help lower the financial barrier for users to check whether they receive the model they pay for.

Cost-efficient model auditing is challenging because reducing generation cost can also remove the statistical signal needed to distinguish models. Under black-box access, responses are stochastic: repeated calls to the same model may differ, while related models can induce highly similar response distributions (Gao et al., 2025; Pasquini et al., 2025; Sun et al., 2025b). Moreover, restricting responses to a small categorical alphabet reduces output cost but may collapse model-specific differences, whereas a larger alphabet may require more samples to estimate reliably (Chan et al., 2014; Canonne, 2020; Weissman et al., 2003). The key question is therefore whether a short output can retain desirable model-specific separation for reliable and inexpensive equality testing.

To address this trade-off, we study a frustratingly simple approach: ask each endpoint to output exactly one character from a fixed alphabet. Rather than compressing arbitrary free-form generations, we directly construct a controlled categorical response. Repeated outputs form an empirical categorical fingerprint, while a paired permutation test accounts for response stochasticity when comparing the candidate and reference models. This design reduces each completion to essentially the minimum generation length, while the explicit alphabet makes the trade-off between model separation and distribution estimation directly measurable. Empirically, across both same-family and cross-family model comparisons, our method achieves substantially lower audit cost than existing black-box baselines at matched detection power and false-positive constraints. Ablation studies further examine how prompt design, alphabet size, and response length affect this cost–power trade-off.

Our **main contributions** are: (i) **Cost-efficient model equality testing.** We formulate black-box LLM equality testing as end-to-end monetary cost minimization under explicit power and false positive constraints; (ii) **A one-character black-box fingerprint.** We introduce a simple categorical fingerprint and a paired permutation test that use only standard text outputs. We further study how alphabet size controls the trade-off between model separation and distribution estimation; and (iii) **Controlled cost and power evaluation.** We establish a common protocol with different-model trials and two same-model controls. Our experiments show substantial cost reductions across both same-family and cross-family comparison.

2 Related Work

Model substitution risks. Recent work studies whether an inference provider delivers the model and computation that it advertises. Model substitution audits test whether a provider replaces an expensive model with a cheaper alternative (Cai et al., 2025b; Sun et al., 2025a; Zhang et al., 2026c). Complementary studies examine quality degradation, resource-consumption threats, and the possibility that users pay premium prices for lower-quality service (Zhang et al., 2026d; Velasco et al., 2026b; Cao et al., 2026). A second line of work audits token accounting and hidden billing behavior through predictive checks or service-side consistency tests (Wang et al., 2025; Sun et al., 2026; Velasco et al., 2026a). These studies establish the economic motivation for remote auditing. Our work supplies a low-cost statistical primitive for the model-identity component of this problem.

Intrusive model fingerprints. Intrusive methods modify training or fine-tuning so that ownership can later be verified. Representative approaches embed backdoor-like behaviors, parameter signatures, or instruction-response mappings into pretrained language models (Gu et al., 2022; Li et al., 2023; Xu et al., 2024). Follow-up work studies fingerprints compatible with parameter-efficient fine-tuning, continual training, model merging, and downstream adaptation (Xu et al., 2025a;b; Yamabe et al., 2025; Tang et al., 2025). Other methods improve scalability or design encrypted model-specific identifiers intended to survive common transformations (Nasery et al., 2025b; Xiong et al., 2026). Semantically conditioned watermarks further broaden the space of model-side ownership mechanisms (Gloaguen et al., 2026). These methods are valuable when the owner can alter the model before deployment. Our setting is non-intrusive: an auditor may receive only query access to an already deployed candidate and a trusted reference.

Non-intrusive white-box model fingerprints. Non-intrusive white-box methods extract model-specific signals without modifying the target model. Early DNN fingerprinting methods use decision boundaries or carefully selected examples to identify a network (Zheng et al., 2022). LLM-oriented approaches inspect parameters, gradients, hidden representations, or attention patterns. EasyDetector (Zhang et al., 2024) detects model reuse from model-level features, HuRef (Zeng et al., 2024) constructs transferable fingerprints from hidden representations, TensorGuard (Wu et al., 2025) studies tensor-based integrity signals, REEF (Zhang et al., 2025) compares invariant representation spaces, and AttnDiff (Zhang et al., 2026a) uses attention differences for fine-grained identification. Recent work also investigates intrinsic lineage and training-stage signals, including SeedPrints (Tong et al., 2026) and LLM DNA (Wu et al., 2026). Such methods can provide rich evidence when the model is locally available, but their internal signals are unavailable for a proprietary API.

Non-intrusive black-box model fingerprints. Non-intrusive black-box methods infer model identity from ordinary input-output behavior. TRAP constructs targeted prompt honeypots, Proflingo profiles responses to linguistic probes, and UTF uses universal transferable fingerprints (Gubri et al., 2024; Jin et al., 2024; Cai et al., 2025a). Other work attributes generated text through stylistic or latent behavioral signatures, including generation fingerprints, palimpsestic membership inference, and chain-of-thought semantic reasoning fingerprints (McGovern et al., 2025; Kuditipudi et al., 2025; Ren et al., 2025). Recent provenance methods exploit internalized knowledge, characteristic error patterns, response-style features, or dataset-specific artifacts (Shao et al., 2026; Zang et al., 2025; Nikolic et al., 2025; Hu et al., 2025; Yan et al., 2026). LLMmap uses a query suite and a classifier to map an endpoint to a known model, while Idiosyncrasies learns an embedding space for closed-set attribution (Pasquini et al., 2025; Sun et al., 2025b). These methods primarily ask *which known model* produced an output. Model equality testing instead asks whether two endpoints induce the same distribution, including when both endpoints come from the same model family.

Within this black-box setting, MET formulates LLM auditing as a two-sample test over prompt-completion pairs and applies maximum mean discrepancy with a Hamming kernel (Gao et al., 2025). RUT derives a rank-based uniformity test, while Token-DiFR studies deterministic inference verification under a stronger decoding assumption (Zhu et al., 2026; Karvonen et al., 2025). ESF and TRAP identify sensitive inputs using model-side signals or optimization during fingerprint construction (Bai et al., 2025; Gubri et al., 2024). B3IT discovers low-temperature border inputs and monitors their first-token output distributions using sampled responses (Chauvin et al., 2026). This family of work is closest to ours because the final decision is a hypothesis test rather than a closed-set classification. Our method differs in using the same fixed one-character channel for both

construction and testing, which removes a separate input-discovery stage and makes the cost of every stage explicit.

Our method is therefore a non-intrusive, strict black-box equality test. It requires neither model modification nor access to weights, representations, gradients, logits, or token probabilities. Unlike closed-set identification, it does not assume that the candidate belongs to a catalog of known models. Unlike construction-heavy change detectors, it does not search for a model-specific prompt before monitoring. Its primary objective is to minimize the complete monetary cost required to attain a fixed detection power under explicit same-model false positive controls.

Adversarial attacks against model fingerprints. Fingerprint reliability must also be evaluated against evasion, removal, and impersonation. Recent studies show that fingerprinting pipelines can be vulnerable to adversarial perturbations and adaptive attacks (Nasery et al., 2025a; Kurian et al., 2025). Other work analyzes the difficulty of identifying a model's origin after fine-tuning or transformation (Yang et al., 2025b). GhostPrint demonstrates a complementary spoofing risk in which an untrusted service can manipulate outputs to imitate a target fingerprint (Zhang et al., 2026b). These threats are orthogonal to the cost question studied here, but they delimit the security guarantee. We evaluate honest but potentially substituted endpoints and do not claim robustness to a fully adaptive provider that knows and actively targets the audit.

3 Problem Formulation

Notations. Let $|r|$ denote the length (i.e., the number of tokens) of a discrete sequence r. We treat an LLM as a conditional probability distribution over token sequences. Given an LLM f parameterized by $\boldsymbol{\theta}$ and a prompt q, a response r is sampled as $r \sim f(\cdot \mid q; \boldsymbol{\theta})$. When the parameters are fixed, we write $r \sim f(\cdot \mid q)$ for notation simplicity.

Black-Box Model Equality Testing (MET). The goal of black-box MET is to audit a third-party LLM API, denoted as f_{api}, to verify whether it matches the advertised reference model, f_{ref}. The auditing procedure relies on statistical hypothesis testing and is formalized below.

Definition 3.1 (Black-Box Model Equality Testing)**.** *Consider an auditor with a query distribution $\mathcal{D}_q$, a discrete query budget n, and a statistical testing mechanism g. The audit consists of the following steps:*

1. ***Query Sampling**: Draw the queries $\mathcal{Q}^n := \{q^i\}_{i=1}^n$ according to $q^i \sim \mathcal{D}_q$.*
2. ***Response Collection**: For each query, obtain responses from both models. Denote the collected generations by $\mathcal{R}^n_{\text{api}} := \{r^i_{\text{api}}\}_{i=1}^n$ and $\mathcal{R}^n_{\text{ref}} := \{r^i_{\text{ref}}\}_{i=1}^n$, with $r^i_{\text{api}} \sim f_{\text{api}}(\cdot \mid q^i)$ and $r^i_{\text{ref}} \sim f_{\text{ref}}(\cdot \mid q^i)$.*
3. ***Verification**: Use g to perform a two-sample test on $\mathcal{R}^n_{\text{api}}$ and $\mathcal{R}^n_{\text{ref}}$. The decision $g(\mathcal{R}^n_{\text{api}}, \mathcal{R}^n_{\text{ref}}) \in \{0, 1\}$ takes the value 0 upon rejection of the null hypothesis and 1 otherwise. The null hypothesis states that the API matches f_{ref}.*

This definition encapsulates general MET frameworks. For instance, Gao et al. Gao et al. (2025) formalize g as a statistical two-sample kernel test (Maximum Mean Discrepancy) that computes the distance between $\mathcal{R}^n_{\text{api}}$ and $\mathcal{R}^n_{\text{ref}}$, thresholding the empirical p-value to yield the binary decision.

Cost-Efficient MET. Auditing commercial APIs incurs financial costs tied to token usage. Let $c_p^{\text{api}}, c_g^{\text{api}}$ and $c_p^{\text{ref}}, c_g^{\text{ref}}$ be the unit prices for prompt (input) and generation (output) tokens for the API and reference model, respectively. The total cost of an audit is:

$$C(\mathcal{Q}^n, \mathcal{R}^n_{\text{api}}, \mathcal{R}^n_{\text{ref}}) = \sum_{i=1}^{n} \left(c_p^{\text{api}} |q^i| + c_g^{\text{api}} |r^i_{\text{api}}| + c_p^{\text{ref}} |q^i| + c_g^{\text{ref}} |r^i_{\text{ref}}| \right).$$

Standard MET requires large query budgets n and long generations $|r^i_{\text{api}}|$ and $|r^i_{\text{ref}}|$ to achieve high statistical power, inflating $C(\mathcal{Q}^n, \mathcal{R}^n_{\text{api}}, \mathcal{R}^n_{\text{ref}})$. *Cost-Efficient MET* aims to construct an optimal query mechanism that minimizes the actual expected cost $\mathbb{E}[C(\mathcal{Q}^n, \mathcal{R}^n_{\text{api}}, \mathcal{R}^n_{\text{ref}})]$ while strictly bounding the test's Type I and Type II error rates.

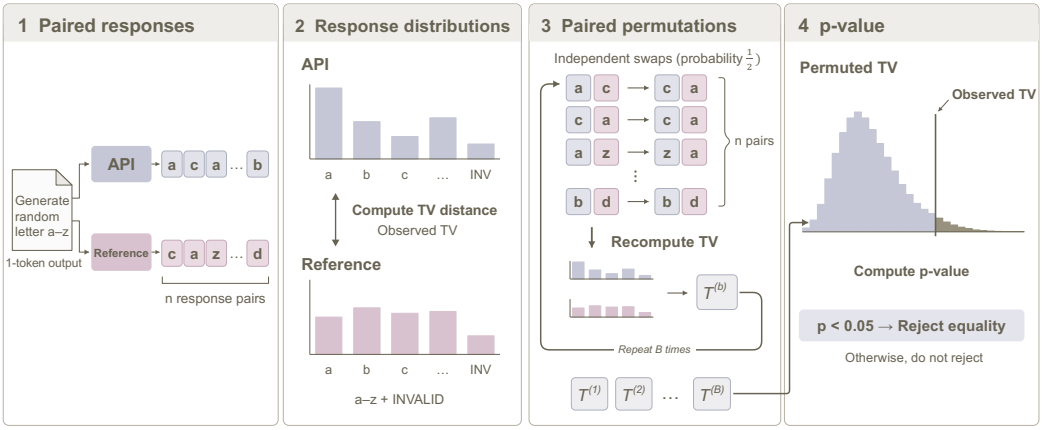


Figure 2: **Overview of the one-character equality test**. We query the API and reference model with the same input and compare their response distributions using total variation (TV) distance. Independent within-pair label swaps give a permutation distribution for computing the p-value.

4 Proposed Method

In this section, we present our one-character black-box fingerprint (see Figure 2), and then present a theoretical result to explain our method.

4.1 The One-Character Fingerprint

One-Character Response. Let $\mathcal{A}$ be a finite output alphabet. Our primary setting uses the 26 lowercase English letters. Each paired observation sends the same query to the API and the reference model. In the primary experiment, the query is the following fixed instruction:

```
Choose a random lowercase English letter from a to z.
Output exactly that one letter and nothing else.
```

The generation is capped at one output token. We apply NFKC normalization, remove surrounding whitespace, and convert the response to lowercase. A response is valid only when the entire normalized string equals one symbol in $\mathcal{A}$. Every other response is mapped to an explicit `invalid` category. We denote the complete categorical output space by $\widetilde{\mathcal{A}} := \mathcal{A} \cup \{\texttt{invalid}\}$.

Let $z^i_{\text{api}}, z^i_{\text{ref}} \in \widetilde{\mathcal{A}}$ denote the parsed responses for the ith paired observation. The resulting samples are $\mathcal{Z}^n_{\text{api}} := \{z^i_{\text{api}}\}_{i=1}^n$ and $\mathcal{Z}^n_{\text{ref}} := \{z^i_{\text{ref}}\}_{i=1}^n$. This construction only constrains the observed response. It does not require the auditor to access the endpoint's tokenizer or token probabilities.

Test Statistic. We compute the empirical categorical distributions

$$\widehat{p}_{\text{api}}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{z^i_{\text{api}} = a\}, \qquad \widehat{p}_{\text{ref}}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{z^i_{\text{ref}} = a\}, \qquad \forall a \in \widetilde{\mathcal{A}},$$

and use their total variation distance as the test statistic

$$T_{\text{obs}} := \text{TV}(\widehat{p}_{\text{api}}, \widehat{p}_{\text{ref}}) = \frac{1}{2}\sum_{a\in\widetilde{\mathcal{A}}} \left|\widehat{p}_{\text{api}}(a) - \widehat{p}_{\text{ref}}(a)\right|.$$

Paired Permutation Test. Under the null hypothesis, the API and reference responses within each paired observation are exchangeable. We use this property to calibrate T_{obs}. For each permutation $b \in \{1, \ldots, B\}$, we independently retain or swap the two endpoint labels within every pair:

$$\left(z^{i,(b)}_{\text{api}}, z^{i,(b)}_{\text{ref}}\right) = \begin{cases} \left(z^i_{\text{api}}, z^i_{\text{ref}}\right), & \xi^{(b)}_i = 0, \\ \left(z^i_{\text{ref}}, z^i_{\text{api}}\right), & \xi^{(b)}_i = 1, \end{cases}$$

where $\xi_i^{(b)} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)$. Let T_b be the TV statistic after the bth permutation. We compute the conservative Monte Carlo p-value

$$\widehat{p} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \geq T_{\text{obs}}\}}{B+1}.$$

At significance level α, the verification mechanism is

$$g\left(\mathcal{Z}_{\text{api}}^n, \mathcal{Z}_{\text{ref}}^n\right) = \mathbf{1}\{\widehat{p} \geq \alpha\}.$$

Thus, $\widehat{p} < \alpha$ rejects model equality. In our empirical study, the formal experiments use $B = 9{,}999$ permutations and $\alpha = 0.05$ as hyper parameters.

4.2 Theoretical Analysis

We study when a single-symbol response retains enough information for model equality testing in the discrete attention model of Ildiz et al. (2024). The result below gives a two-token query and a sample bound for the exact paired permutation test.

For an integer $K \geq 2$, let $[K] := \{1, \ldots, K\}$, and let $\mathbf{I}_K$ denote the $K \times K$ identity matrix. All logarithms are natural. Write $\text{Bern}(p)$ for the distribution on $\{0, 1\}$ with probability p at one. For $t \in \mathbb{R}$ and $B \geq 0$, define $\sigma(t) := (1 + e^{-t})^{-1}, \kappa(B) := \frac{e^B}{(1+e^B)^2}$. Then, we have

Definition 4.1 (Attention response distribution)**.** *Let $\mathbf{A} \in \mathbb{R}^{K \times K}$. A query is a finite sequence $q = (a_1, \ldots, a_L) \in [K]^L$, where $L \geq 1$. Define $k := a_L$ and $c_i(q) := |\{t \in [L] : a_t = i\}|$ for $i \in [K]$. The model returns one token $Y \in [K]$ with probabilities*

$$p_{\mathbf{A}}(i \mid q) := \frac{c_i(q) e^{A_{ik}}}{\sum_{j=1}^{K} c_j(q) e^{A_{jk}}}.$$

Let $\mathbf{p}_{\mathbf{A}}(q) := (p_{\mathbf{A}}(i \mid q))_{i=1}^K$. Each call uses fresh randomness, independently of previous calls conditional on the selected endpoint and query.

This is the original attention model written in terms of its token scores. Specifically, let $d \geq K$ be an integer, let $\mathbf{E} \in \mathbb{R}^{K \times d}$ have linearly independent rows, let $\mathbf{C} \in \mathbb{R}^{K \times d}$ satisfy $\mathbf{C}\mathbf{E}^\top = \mathbf{I}_K$, and let $\mathbf{W} \in \mathbb{R}^{d \times d}$ be the attention matrix. Grouping repeated input tokens gives Definition 4.1 with $\mathbf{A} = \mathbf{E}\mathbf{W}\mathbf{E}^\top$. The model has no positional embeddings, residual connection, layer normalization, or MLP. It samples from the displayed probabilities without a second vocabulary softmax.

Definition 4.2 (Differences between two models)**.** *Let $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{R}^{K \times K}$ specify two endpoints. For $m \in \{0, 1\}$ and $i, k \in [K]$, define*

$$\begin{aligned} b_m(i,k) &:= (A_m)_{ik} - (A_m)_{kk}, \\ \Delta &:= \max_{i \neq k} |b_1(i,k) - b_0(i,k)|, \\ B &:= \max_{m \in \{0,1\},\, i,k \in [K]} |b_m(i,k)|. \end{aligned}$$

The equality null requires $\mathbf{p}_{\mathbf{A}_0}(q) = \mathbf{p}_{\mathbf{A}_1}(q)$ for every finite nonempty query q.

Definition 4.3 (Permutation test for two possible responses)**.** *Fix a query with two possible responses, encoded as zero and one. Let $n \geq 1$ be an integer and let $p_0, p_1 \in [0, 1]$. For each endpoint $m \in \{0, 1\}$, observe n independent samples $X_{m,1}, \ldots, X_{m,n} \sim \text{Bern}(p_m)$. Assume independence between endpoints as well. Define $Z_j := X_{0,j} - X_{1,j}, T := \frac{1}{n}\left|\sum_{j=1}^n Z_j\right|$.*

Let $\xi_1, \ldots, \xi_n$ be independent uniform signs in $\{-1, 1\}$, independent of the data. The exact permutation p-value is

$$\pi := \Pr_{\xi}\left[\left|\sum_{j=1}^{n} \xi_j Z_j\right| \geq \left|\sum_{j=1}^{n} Z_j\right| \,\middle|\, Z_1, \ldots, Z_n\right].$$

At level $\eta \in (0, 1)$, reject equality if $\pi \leq \eta$.

Multiplication by $\xi_j = -1$ swaps the two endpoint responses in pair j. The definition includes ties in the p-value.

Theorem 4.4 (Sample complexity for a two-token query)**.** *Under Definitions 4.1–4.2, the endpoint distributions agree for every query if and only if $\Delta = 0$. Suppose $\Delta > 0$ and choose $(i_\star, k_\star) \in \arg\max_{i \neq k} |b_1(i,k) - b_0(i,k)|$. Let $q_\star := (i_\star, k_\star)$ and encode its responses $i_\star, k_\star$ as one and zero. For $0 < \eta \leq 1/3$, the test in Definition 4.3 has type-I error at most η and power at least $1 - \eta$ using*

$$n = O\left(\frac{1}{\kappa(B)\Delta^2} \log \frac{1}{\eta}\right)$$

samples from each endpoint at $q_\star$. It uses $2n$ calls, $4n$ input tokens, and $2n$ output tokens. The rejection rule does not use B, Δ, or the response probabilities.

The hidden constant is universal. For bounded B, the sample bound is $O(\Delta^{-2} \log(1/\eta))$, with no explicit dependence on K. The query is supplied to the test, so the bound excludes query selection. Our LLM experiments use the instruction in Section 4.1 and Monte Carlo permutation sampling. The proof is in Appendix A.

5 Experiments

We evaluate the cost efficiency of one-character model equality testing. We compare our method with existing black-box tests on within-family and cross-family model substitutions, then study the effects of the prompt, alphabet size, response length, and alphabet choice.

5.1 Experimental Setup

Baselines. We compare with MET (Gao et al., 2025), which uses a Hamming-kernel MMD two-sample test, and B3IT (Chauvin et al., 2026), which identifies border inputs and compares their first-token responses. Both methods audit the target endpoint through sampled outputs.

Evaluation protocol and power. Following MET (Gao et al., 2025), we evaluate detection power at a fixed nominal significance level $\alpha = 0.05$. Power is the probability that a test rejects model equality when the two models differ. For each configuration, we estimate power as the fraction of $R = 25$ repeated audits that reject equality on a different-model pair. This protocol measures detection reliability while accounting for randomness in model outputs. Shading indicates pointwise 95% Wilson confidence intervals. Same-model false-positive-rate results are reported in Appendix B.

Cost and comparison criterion. Cost is the mean monetary expenditure per audit in USD, averaged over the $R = 25$ repetitions. It includes input- and output-token charges for both endpoints, together with border discovery and calibration for B3IT. We use monetary cost to account for differences in query length, response length, and endpoint pricing. In each cost–power plot, the horizontal axis shows mean audit cost on a logarithmic scale, and the vertical axis shows empirical power. Higher power at a fixed cost, or lower cost at a fixed power, indicates greater efficiency.

Query construction. Our primary setting repeats the following instruction: "`Choose a random lowercase English letter from a to z. Output exactly that one letter and nothing else.`" The alphabet is $\mathcal{A}_{26} = \{\mathsf{a}, \ldots, \mathsf{z}\}$, and each generation is capped at one output token. We apply NFKC normalization, strip surrounding whitespace, and lowercase the response. Any response outside the allowed alphabet is assigned to an explicit INVALID category. Each paired observation contains one completion from each endpoint under the same query, so a budget of n provides n responses from each endpoint. We compare the empirical categorical distributions using TV distance and calibrate the test with $9{,}999$ within-pair permutations.

MET samples user messages from the `test_sft` split of `HuggingFaceH4/ultrachat_200k`. Our method and MET use temperature 1.0 and top-p 1.0. We fix the provider route for each endpoint and record realized token usage and charges. Provider, precision, and collection time can also affect comparisons between hosted endpoints.

5.2 Equality-Test Performance

We evaluate four within-family size changes and four cross-family substitutions involving Llama 3.1, Qwen3, Ministral 3, and Gemma 3. Figures 3 and 4 show the corresponding cost–power comparisons. For our method, we evaluate query budgets $n \in \{1, \ldots, 100, 150, 200, 300\}$ per endpoint. MET

varies the number of prompts over $m \in \{5, 10, 25\}$, the output-token cap over $L \in \{5, 10, 25, 50\}$, and the sample budget. Each MET response independently samples a prompt uniformly with replacement. We apply Hamming MMD to Unicode-codepoint sequences using 999 global row permutations. B3IT varies the number of border inputs and detection samples per input. Responses are reused across nested query budgets.

For Ours and MET, we plot the empirical cost–power Pareto frontiers over the evaluated parameter grid. We retain configurations for which no evaluated setting has lower cost and at least as much power, or higher power at no greater cost. For B3IT, we show the configuration with the highest empirical power, breaking ties by cost.

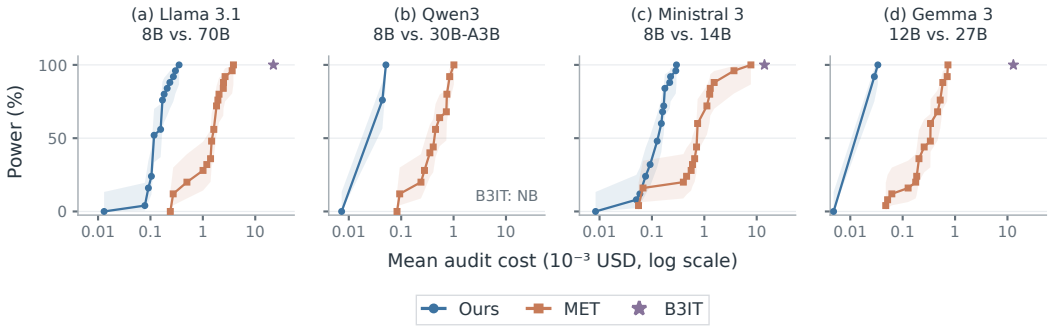


Figure 3: Cost–power comparison for four within-family model pairs. Ours and MET show empirical Pareto frontiers, and B3IT shows its highest-power configuration. All methods use $R = 25$ audits. Shading gives pointwise 95% Wilson intervals, unadjusted for configuration selection. NB denotes no validated B3IT borders.

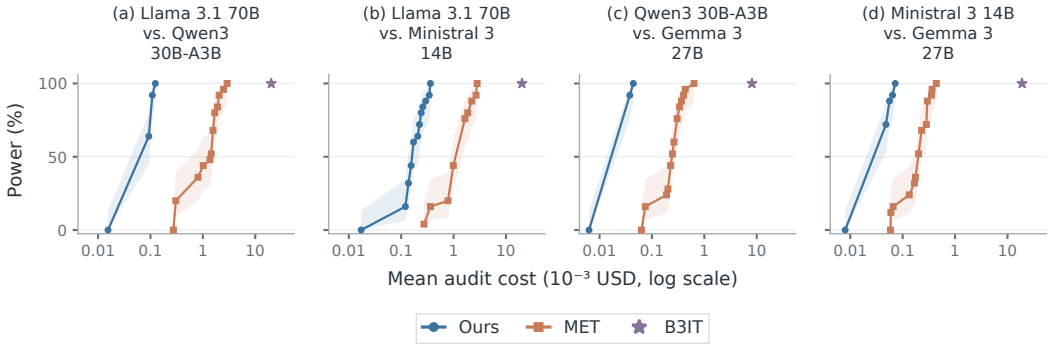


Figure 4: Cost–power comparison for four cross-family model pairs. The evaluation and configuration-selection rules follow Figure 3. Both figures use the same logarithmic cost scale.

Figures 3 and 4 show that our method reaches high empirical power at lower observed cost across the evaluated model pairs. Compared with MET, our curves reach the same high-power region with substantially less expenditure. The reported B3IT configurations also incur higher total cost when border discovery and calibration are included. The advantage holds for both within-family size changes and cross-family substitutions, suggesting that the one-character response retains useful model-specific differences across these settings. These results support short categorical responses as a cost-efficient choice for repeated model auditing.

5.3 Ablation Study

We use three representative model pairs for all ablations: Llama-3.1-70B-Instruct paired with Llama-3.1-8B-Instruct, Qwen3-30B-A3B-Instruct, or Hermes-3-Llama-3.1-70B. These cover a size change, a cross-family substitution, and a fine-tuned derivative. The panels follow this order in every ablation figure. All ablations use the same TV statistic, 9,999 paired permutations, nominal significance level $\alpha = 0.05$, and $R = 25$ audits per configuration. Each curve shows all evaluated query budgets, with cost and pointwise confidence intervals computed as in the main comparison. Additional ablations on alphabet size and response length are reported in Appendix C.

Impact of the prompt. We compare the instruction alone with an UltraChat user message followed by the same instruction. The latter condition uses messages from the `test_sft` split, with the same message sent to both endpoints. Both conditions request one lowercase letter from $\mathcal{A}_{26}$ and use the same response parser. This comparison examines if adding language context improves efficiency.

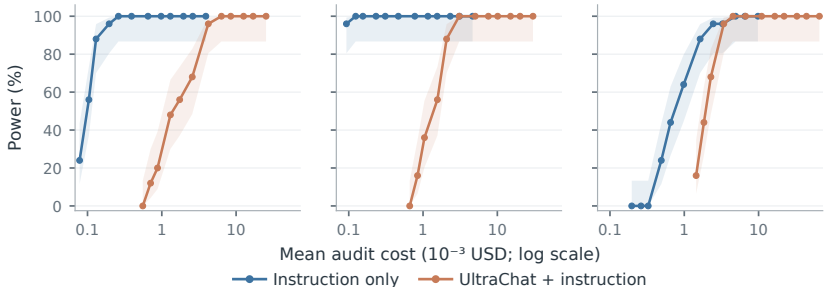


Figure 5: Prompt ablation. We compare the instruction alone with an UltraChat user message followed by the instruction. Panels show the size-change, cross-family, and fine-tuning comparisons from left to right.

Figure 5 shows that the instruction alone reaches high power at lower cost in all three comparisons. The advantage is particularly clear for the size-change and cross-family pairs. Although both prompt variants require more expenditure to distinguish the fine-tuned derivative, the instruction-only query remains more efficient. The additional UltraChat context increases input cost without improving the observed cost–power trade-off, supporting a short, context-free instruction as a practical default.

Impact of alphabet choice. At $K = 10$, we compare lowercase letters $\{\mathsf{a}, \ldots, \mathsf{j}\}$, uppercase letters $\{\mathsf{A}, \ldots, \mathsf{J}\}$, and digits $\{0, \ldots, 9\}$. We fix the prompt structure, response length, and statistical test, and disable case folding for the uppercase condition. Holding the alphabet size fixed isolates the effect of symbol choice.

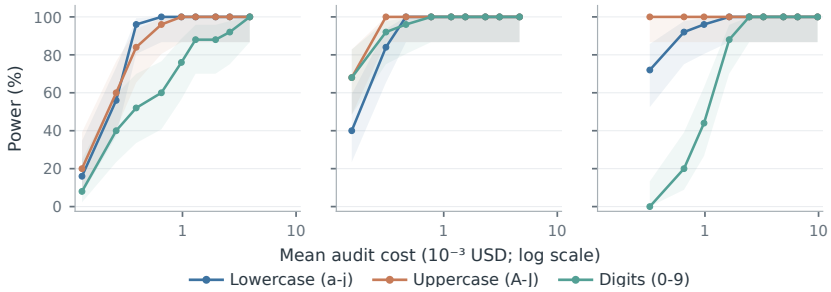


Figure 6: Alphabet-choice ablation. Each query requests one character from lowercase a–j, uppercase A–J, or digits 0–9, with $K = 10$.

Figure 6 shows that equal-sized alphabets can produce different cost–power curves. Lowercase letters perform well for the size-change comparison, while uppercase letters are particularly effective for the fine-tuned derivative. Digits require more expenditure to reach high power in these two comparisons. Thus, alphabet size alone does not determine auditing efficiency. The choice of symbols also influences how clearly the response distributions distinguish the models.

6 Conclusion

We study cost-efficient black-box model equality testing and propose a simple approach that requests one character from each model per response. Repeated responses form categorical distributions, which we compare using total variation distance and a paired permutation test. Experiments across same-family and cross-family model pairs demonstrate favorable cost–power trade-offs compared with existing methods. Ablation studies further show how prompt design, alphabet choice, and response length affect auditing efficiency. Our theoretical analysis establishes that, in a discrete attention model, a two-token query with single-symbol responses suffices to distinguish models with different response distributions.

AI USE STATEMENT

In this work, we used generative AI tools for tasks requiring disclosure, including assistance with writing proofs, and for tasks for which disclosure is recommended, including suggesting experimental parameters, creating or editing software code, improving the readability of the manuscript, and proposing a title or keywords. We have reviewed all AI-assisted work. All code has been reviewed by a human, and every proof step and mathematical statement has been double-checked by at least one coauthor. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Xiaofan Bai, Pingyi Hu, Xiaojing Ma, Linchen Yu, Dongmei Zhang, Qi Zhang, and Bin Benjamin Zhu. Esf: Efficient sensitive fingerprinting for black-box tamper detection of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10477–10494, 2025.

Jiacheng Cai, Jiahao Yu, Yangguang Shao, Yuhang Wu, and Xinyu Xing. UTF: Under-trained tokens as fingerprints: A novel approach to LLM identification. In *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pages 1–6, 2025a.

Will Cai, Tianneng Shi, Xuandong Zhao, and Dawn Song. Are you getting what you pay for? auditing model substitution in llm apis. *arXiv preprint arXiv:2504.04715*, 2025b.

Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing Graduate Surveys*, 9:1–100, 2020. doi: 10.4086/toc.gs.2020.009.

Yuhan Cao, Yu Wang, Sitong Liu, Miao Li, Yixin Tao, and Tianxing He. Pay for the second-best service: A game-theoretic approach against dishonest llm providers. In *Proceedings of the ACM Web Conference 2026*, pages 249–260, 2026.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 1193–1203. SIAM, 2014.

Timothée Chauvin, Clément Lalanne, Erwan Le Merrer, Jean-Michel Loubes, François Taïani, and Gilles Tredan. Token-efficient change detection in llm apis. In *ICML*, 2026.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *International Conference on Machine Learning*, pages 8359–8388. PMLR, 2024.

Irena Gao, Percy Liang, and Carlos Guestrin. Model equality testing: Which model is this api serving? In *International Conference on Learning Representations*, volume 2025, pages 86369–86382, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL `https://arxiv.org/abs/2503.19786`.

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. LLM fingerprinting via semantically conditioned watermarks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Chenxi Gu, Chengsong Huang, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Watermarking pre-trained language models with backdooring. *arXiv preprint arXiv:2210.07543*, 2022.

Martin Gubri, Dennis Ulmer, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Trap: Targeted random adversarial prompt honeypot for black-box identification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11496–11517, 2024.

Pingyi Hu, Xiaofan Bai, Xiaojing Ma, Chaoxiang He, Dongmei Zhang, and Bin Benjamin Zhu. Resf: Regularized-entropy-sensitive fingerprinting for black-box tamper detection of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4889–4903, 2025.

Muhammed Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to Markov models: Unveiling the dynamics of generative transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20955–20982. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ildiz24a.html.

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

Adam Karvonen, Daniel Reuter, Roy Rinberg, Luke Marks, Adrià Garriga-Alonso, and Keri Warr. Difr: Inference verification despite nondeterminism. *arXiv preprint arXiv:2511.20621*, 2025.

Rohith Kuditipudi, Jing Huang, Sally Zhu, Diyi Yang, Christopher Potts, and Percy Liang. Blackbox model provenance via palimpsestic membership inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Kevin Kurian, Ethan Holland, and Sean Oesch. Attacks and defenses against llm fingerprinting. *arXiv preprint arXiv:2508.09021*, 2025.

Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–34, 2022.

Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. Plmmark: a secure and robust black-box watermarking framework for pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14991–14999, 2023.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui

Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. Your large language models are leaving fingerprints. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 85–95, 2025.

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

Anshul Nasery, Jonathan Hayase, Creston Brooks, Peiyao Sheng, Himanshu Tyagi, Pramod Viswanath, and Sewoong Oh. Scalable fingerprinting of large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

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*, 2025.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. LLMmap: Fingerprinting for large language models. In *USENIX Security Symposium*, 2025.

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.

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

Shuo Shao, Yiming Li, Hongwei Yao, Yifei Chen, Yuchen Yang, and Zhan Qin. Reading between the lines: Towards reliable black-box llm fingerprinting via zeroth-order gradient estimation. In *Proceedings of the ACM Web Conference 2026*, pages 2637–2648, 2026.

Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–29, 2021.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Guoheng Sun, Ziyao Wang, Bowei Tian, Meng Liu, Zheyu Shen, Shwai He, Yexiao He, Wanghao Ye, Yiting Wang, and Ang Li. Coin: Counting the invisible reasoning tokens in commercial opaque llm apis. *arXiv preprint arXiv:2505.13778*, 2025a.

Guoheng Sun, Ziyao Wang, Xuandong Zhao, Bowei Tian, Zheyu Shen, Yexiao He, Jinming Xing, and Ang Li. Position: Invisible tokens, visible bills: The urgent need to audit hidden operations in opaque llm services. In *Forty-third International Conference on Machine Learning (ICML) Position Paper Track*, 2026.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *International Conference on Machine Learning*, pages 57854–57885. PMLR, 2025b.

Ling Tang, YueFeng Chen, Hui Xue, and Quanshi Zhang. Towards the resistance of neural network fingerprinting to fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yao Tong, Haonan Wang, Siquan Li, Kenji Kawaguchi, and Tianyang Hu. Seedprints: Fingerprints can even tell which seed your large language model was trained from. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ander Artola Velasco, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Auditing pay-per-token in large language models. In *AISTATS*, 2026a.

Ander Artola Velasco, Stratis Tsirtsis, Nastaran Okati, and Manuel Gomez Rodriguez. Is your LLM overcharging you? tokenization, transparency, and incentives. In *Forty-third International Conference on Machine Learning (ICML)*, 2026b.

Ziyao Wang, Guoheng Sun, Yexiao He, Zheyu Shen, Bowei Tian, and Ang Li. Predictive auditing of hidden tokens in llm apis via reasoning length estimation. *arXiv preprint arXiv:2508.00912*, 2025.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97 (R.1), Hewlett-Packard Laboratories, 2003.

Zehao Wu, Yanjie Zhao, and Haoyu Wang. Tensorguard: Gradient-based model fingerprinting for llm similarity detection and family classification. In *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 445–456. IEEE, 2025.

Zhaomin Wu, Haodong Zhao, Ziyang Wang, Jizhou Guo, Qian Wang, and Bingsheng He. LLM DNA: Tracing model evolution via functional representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zixun Xiong, Gaoyi Wu, Qingyang Yu, Mingyu Derek Ma, Lingfeng Yao, Miao Pan, Xiaojiang Du, and Hao Wang. iseal: Encrypted fingerprinting for reliable llm ownership verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40:42, pages 35940–35948, 2026.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3277–3306, 2024.

Zhenhua Xu, Zhaokun Yan, Binhan Xu, Xin Tong, Haitao Xu, Yourong Chen, and Meng Han. Unlocking the effectiveness of lora-fp for seamless transfer implantation of fingerprints in downstream models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4302–4312, 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*, pages 6978–7000, 2025b.

Shojiro Yamabe, Futa Kai Waseda, Tsubasa Takahashi, and Koki Wataoka. Mergeprint: Merge-resistant fingerprints for robust black-box ownership verification of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6894–6916, 2025.

Yuliang Yan, Haochun Tang, Shuo Yan, and Enyan Dai. Duffin: A dual-level fingerprinting framework for llms ip protection. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5168–5184, 2026.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL `https://arxiv.org/abs/2505.09388`.

Ziqing Yang, Yixin Wu, Yun Shen, Wei Dai, Michael Backes, and Yang Zhang. The challenge of identifying the origin of black-box large language models. *arXiv preprint arXiv:2503.04332*, 2025b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Chuanchao Zang, Xiangtao Meng, Wenyu Chen, Tianshuo Cong, Zha Yaxing, Dong Qi, Zheng Li, and Shanqing Guo. Errortrace: A black-box traceability mechanism based on model family error space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. Huref: Human-readable fingerprint for large language models. *Advances in Neural Information Processing Systems*, 37:126332–126362, 2024.

Haobo Zhang, Zhenhua Xu, Junxian Li, Shangfeng Sheng, Dezhang Kong, and Meng Han. Attndiff: Attention-based differential fingerprinting for large language models. In *ACL*, 2026a.

Jiahao Zhang, Xiuyu Li, and Suhang Wang. Your Pro LLM subscription may actually be Free: Exposing fingerprint spoofing risks in LLM inference services. *arXiv preprint arXiv:2606.16100*, 2026b. URL https://arxiv.org/abs/2606.16100.

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

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation encoding fingerprints for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yage Zhang, Yukun Jiang, Zeyuan Chen, Michael Backes, Xinyue Shen, and Yang Zhang. Real money, fake models: Deceptive model claims in shadow APIs. *arXiv preprint arXiv:2603.01919*, 2026c.

Yuanhe Zhang, Xinyue Wang, Zhican Chen, Weiliu Wang, Zilu Zhang, Zhengshuo Gong, Zhenhong Zhou, Kun Wang, Li Sun, Yang Liu, et al. Resource consumption threats in large language models. *arXiv preprint arXiv:2603.16068*, 2026d.

Yue Zheng, Si Wang, and Chip-Hong Chang. A dnn fingerprint for non-repudiable model ownership identification and piracy detection. *IEEE Transactions on Information Forensics and Security*, 17: 2977–2989, 2022.

Xiaoyuan Zhu, Yaowen Ye, Tianyi Alex Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. Auditing black-box LLM APIs with a rank-based uniformity test. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PSIe9mmF7a.

Appendix

LIST OF CONTENTS

In this appendix, we provide the following additional information:

- **Section A**: Proof of the Discrete-Attention Result.
- **Section B**: Same-Model False-Positive Rates.
- **Section C**: Additional Ablation Studies.

A PROOF OF THE DISCRETE-ATTENTION RESULT

We prove Theorem 4.4 using the following lemmas. For an event A, let $\mathbf{1}\{A\}$ be its indicator. For an integer $n \geq 0$ and $p \in [0,1]$, $\mathrm{Bin}(n,p)$ is the law of the sum of n independent $\mathrm{Bern}(p)$ variables.

Lemma A.1 (Response formula and equality). *Let $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{R}^{K\times K}$ satisfy Definitions 4.1 and 4.2. For distinct $i,k \in [K]$ and $m \in \{0,1\}$, the query $q = (i,k)$ satisfies*

$$\begin{aligned} p_{\mathbf{A}_m}(i \mid q) &= \sigma(b_m(i,k)), \\ p_{\mathbf{A}_m}(k \mid q) &= 1 - \sigma(b_m(i,k)), \\ p_{\mathbf{A}_m}(j \mid q) &= 0 \qquad (j \notin \{i,k\}). \end{aligned}$$

Moreover, $\Delta = 0$ if and only if $\mathbf{p}_{\mathbf{A}_0}(q) = \mathbf{p}_{\mathbf{A}_1}(q)$ for every finite nonempty query q.

Proof. For $q = (i,k)$, the counts in Definition 4.1 are $c_i(q) = c_k(q) = 1$ and $c_j(q) = 0$ for $j \notin \{i,k\}$. The last symbol is k. Substituting these counts gives

$$\begin{aligned} p_{\mathbf{A}_m}(i \mid q) &= \frac{e^{(A_m)_{ik}}}{e^{(A_m)_{ik}} + e^{(A_m)_{kk}}} \\ &= \frac{1}{1 + e^{-((A_m)_{ik} - (A_m)_{kk})}} \\ &= \sigma(b_m(i,k)). \end{aligned}$$

The second equality divides the numerator and denominator by $e^{(A_m)_{ik}}$. The final equality uses the definitions of b_m and σ. Only i and k have positive counts, which proves the remaining response formulas.

Suppose the response laws agree for every query. Apply this hypothesis to every query (i,k) with $i \neq k$. The first formula gives $\sigma(b_0(i,k)) = \sigma(b_1(i,k))$. The function σ is strictly increasing, so $b_0(i,k) = b_1(i,k)$ for all distinct i,k. Thus $\Delta = 0$.

Conversely, suppose $\Delta = 0$. Fix a finite nonempty query q ending in k, and define $\lambda_k := (A_1)_{kk} - (A_0)_{kk}$. For $j \neq k$, equality of $b_0(j,k)$ and $b_1(j,k)$ gives $(A_1)_{jk} = (A_0)_{jk} + \lambda_k$. The same identity holds for $j = k$ by the definition of λ_k. Hence, for every $i \in [K]$,

$$\begin{aligned} p_{\mathbf{A}_1}(i \mid q) &= \frac{c_i(q) e^{(A_0)_{ik} + \lambda_k}}{\sum_{j=1}^{K} c_j(q) e^{(A_0)_{jk} + \lambda_k}} \\ &= \frac{c_i(q) e^{(A_0)_{ik}}}{\sum_{j=1}^{K} c_j(q) e^{(A_0)_{jk}}} \\ &= p_{\mathbf{A}_0}(i \mid q). \end{aligned}$$

The second equality cancels the positive factor e^{λ_k}. The denominator is positive because the query is nonempty. This proves the converse. □

For the following lemmas, use the binary observations in Definition 4.3. We write them explicitly to fix the notation. For an integer $n \geq 1$, let the variables $X_{m,j} \sim \mathrm{Bern}(p_m)$, $m \in \{0,1\}$ and $j \in [n]$,

be mutually independent. Define

$$Z_j := X_{0,j} - X_{1,j}, \qquad S := \sum_{j=1}^{n} Z_j, \qquad M := \sum_{j=1}^{n} |Z_j|,$$
$$g := |p_0 - p_1|, \qquad r := p_0(1-p_1) + p_1(1-p_0).$$

Thus M is the number of pairs whose responses differ, and r is the probability that the responses in one pair differ. The empirical total variation statistic is $|S|/n$. For independent uniform signs $\xi_j \in \{-1, 1\}$, its exact paired permutation p-value is

$$\pi := \Pr_{\xi}\left[\left|\sum_{j=1}^{n} \xi_j Z_j\right| \geq |S| \,\middle|\, Z_1, \ldots, Z_n\right].$$

When $M = 0$, this definition gives $\pi = 1$.

Lemma A.2 (Distribution of the pairs whose responses differ)**.** *The count M has distribution $\mathrm{Bin}(n, r)$. If $r > 0$, conditional on $M = m$, the m nonzero values of Z_j are independent signs, each with probability $p_0(1-p_1)/r$ of being positive. Conditional on any observed data with $M \geq 1$, the permutation p-value satisfies*

$$\pi \leq 2\exp\left(-\frac{S^2}{2M}\right).$$

Proof. Independence of the two responses in one pair gives

$$\Pr[Z_j = 1] = p_0(1-p_1),$$
$$\Pr[Z_j = -1] = p_1(1-p_0),$$
$$\Pr[|Z_j| = 1] = r.$$

Independence across pairs proves the binomial law of M. Conditional on any specified set of nonzero indices, the corresponding signs remain independent, each positive with probability $p_0(1-p_1)/r$. This product distribution depends on the number of selected indices but not their identities. Conditioning only on $M = m$ therefore gives the same distribution for the nonzero signs in index order.

Next condition on the observed values of $Z_1, \ldots, Z_n$. Multiplying each nonzero Z_j by an independent uniform sign produces an independent uniform sign. The zero values remain zero. Thus $\sum_j \xi_j Z_j$ has the same distribution as the sum of M independent uniform signs. Hoeffding's two-sided inequality at threshold $|S|$ gives the stated bound. □

Lemma A.3 (Type-I error of the exact test)**.** *For every $p_0 = p_1 \in [0, 1]$ and $\alpha \in (0, 1)$,*

$$\Pr[\pi \leq \alpha] \leq \alpha.$$

Proof. Condition on the unordered responses in each pair. When the two responses differ, their two possible orders have equal probability under $p_0 = p_1$. These orders are independent across pairs conditional on the unordered responses. Consequently, the conditional distribution of the observed statistic agrees with its permutation distribution.

To account for ties, let V be any random variable with finitely many real values and let V' be an independent copy. For a possible value v, define $a(v) := \Pr[V' \geq v]$. The possible values satisfying $a(v) \leq \alpha$ form an upper tail whose total probability is at most α. Therefore $\Pr[a(V) \leq \alpha] \leq \alpha$. Apply this fact to the conditional distribution of $|S|/n$. When $M = 0$, the statistic is zero and $\pi = 1$, so the same conclusion holds. Averaging over the unordered responses proves the claim. □

Lemma A.4 (Error bounds for the exact test)**.** *For $\alpha \in (0, 1)$, the exact test has type-I error at most α under every $p_0 = p_1$. Now let $p_0 \neq p_1$ and let $\alpha, \beta \in (0, 1)$. If*

$$n \geq \frac{16r}{g^2} \log \frac{2}{\min\{\alpha, \beta\}},$$

then $\Pr[\pi \le \alpha] \ge 1 - \beta$.

Proof. The type-I error statement is Lemma A.3. We prove the power statement. Exchanging the endpoint names changes the sign of S but not the test. We may therefore assume $p_0 > p_1$. Since $g > 0$, also $r > 0$. Define $\theta := g/r$. The probabilities of $Z_j = 1$ and $Z_j = -1$ sum to r and differ by g. Thus $0 < \theta \le 1$. By Lemma A.2,

$$\mathbb{E}[S \mid M = m] = m\theta.$$

Define

$$\mathcal{E} := \{M \ge nr/2\} \cap \{S \ge M\theta/2\}.$$

The lower-tail Chernoff bound for $M \sim \mathrm{Bin}(n, r)$ gives

$$\Pr[M < nr/2] \le e^{-nr/8}.$$

For every integer $m \ge nr/2$, Hoeffding's inequality conditional on $M = m$ gives

$$\begin{aligned}\Pr[S < m\theta/2 \mid M = m] &\le e^{-m\theta^2/8} \\ &\le \exp\left(-\frac{ng^2}{16r}\right).\end{aligned}$$

The second inequality uses $m \ge nr/2$ and $\theta = g/r$. Since $g \le r$, a union bound gives

$$\begin{aligned}\Pr[\mathcal{E}^{\mathsf{c}}] &\le e^{-nr/8} + \exp\left(-\frac{ng^2}{16r}\right) \\ &\le 2\exp\left(-\frac{ng^2}{16r}\right) \\ &\le \beta.\end{aligned}$$

The last inequality follows from the assumed sample size.

On $\mathcal{E}$, $M \ge nr/2 > 0$. Lemma A.2 gives

$$\begin{aligned}\pi &\le 2\exp\left(-\frac{S^2}{2M}\right) \\ &\le 2e^{-M\theta^2/8} \\ &\le 2\exp\left(-\frac{ng^2}{16r}\right) \\ &\le \alpha.\end{aligned}$$

The second inequality uses $S \ge M\theta/2$. The third uses $M \ge nr/2$. Thus the test rejects on an event of probability at least $1 - \beta$. □

Lemma A.5 (A bound for two logistic probabilities)**.** *Let $B \ge 0$ and let $x, y \in [-B, B]$ with $x \ne y$. Define $\delta := |x - y|$, $p_0 := \sigma(x)$, $p_1 := \sigma(y)$, $g := |p_0 - p_1|$, and $r := p_0(1 - p_1) + p_1(1 - p_0)$. Then*

$$\frac{r}{g^2} \le \frac{4}{\kappa(B)\delta^2}.$$

Proof. Define $a := (x + y)/2$ and $h := |x - y|/2$. Then $h > 0$ and $|a| + h = \max\{|x|, |y|\} \le B$. For $t \in \mathbb{R}$, define $\cosh(t) := (e^t + e^{-t})/2$ and $\sinh(t) := (e^t - e^{-t})/2$. Substituting p_0 and p_1 gives

$$\begin{aligned}g &= \frac{|e^x - e^y|}{(1 + e^x)(1 + e^y)} = \frac{\sinh(h)}{\cosh(a) + \cosh(h)}, \\ r &= \frac{e^x + e^y}{(1 + e^x)(1 + e^y)} = \frac{\cosh(h)}{\cosh(a) + \cosh(h)}.\end{aligned}$$

For each second equality, divide the numerator and denominator by $2e^a$.

For nonnegative u, v, the addition formula for $\cosh$ gives

$$\cosh(u+v) + 1 - \cosh(u) - \cosh(v) = (\cosh(u) - 1)(\cosh(v) - 1) + \sinh(u)\sinh(v) \geq 0.$$

Taking $u = |a|$ and $v = h$ gives

$$\cosh(a) + \cosh(h) \leq 1 + \cosh(B).$$

This uses $|a| + h \leq B$ and monotonicity of $\cosh$ on $[0, \infty)$. Since $\kappa(B) = 1/(2 + 2\cosh(B))$ and $\delta = 2h$, we obtain

$$\begin{aligned}
\kappa(B)\delta^2 \frac{r}{g^2} &= \frac{2h^2 \cosh(h)(\cosh(a) + \cosh(h))}{(1 + \cosh(B))\sinh^2(h)} \\
&\leq \frac{2h^2 \cosh(h)}{\sinh^2(h)} \\
&\leq \frac{4(\cosh(h) - 1)\cosh(h)}{\sinh^2(h)} \\
&= \frac{4\cosh(h)}{\cosh(h) + 1} \\
&\leq 4.
\end{aligned}$$

The second inequality uses $h^2 \leq 2(\cosh(h) - 1)$, which follows from the power series for $\cosh$. The following equality uses $\sinh^2(h) = (\cosh(h) - 1)(\cosh(h) + 1)$. This proves the result. □

Proof of Theorem 4.4. The equivalence between $\Delta = 0$ and equality at every query is Lemma A.1. Since K is finite and $\Delta > 0$, choose distinct $i_\star, k_\star \in [K]$ that attain the maximum defining Δ. Let $q_\star := (i_\star, k_\star)$. By Lemma A.1, both response distributions are supported on $\{i_\star, k_\star\}$. Encode response $i_\star$ as one and response $k_\star$ as zero. The two Bernoulli parameters are

$$p_0 := \sigma(b_0(i_\star, k_\star)), \qquad p_1 := \sigma(b_1(i_\star, k_\star)).$$

Their two logistic arguments lie in $[-B, B]$ and differ by Δ. They are distinct because σ is strictly increasing. Lemma A.5 therefore gives

$$\frac{r}{g^2} \leq \frac{4}{\kappa(B)\Delta^2}.$$

Substituting this bound in Lemma A.4 shows that

$$n = \left\lceil \frac{64}{\kappa(B)\Delta^2} \log \frac{2}{\eta} \right\rceil$$

samples per endpoint suffice for type-II error at most η. Lemma A.3 gives type-I error at most η for every equal pair of binary response laws at this query. The test compares the exact permutation p-value with η and does not use g, r, B, or Δ in its rejection rule. These quantities only enter the sufficient sample-size bound. Since $r/g^2 \geq 1$, the ratio bound above implies $\kappa(B)\Delta^2 \leq 4$. Thus the ceiling can be absorbed into a universal constant in the stated asymptotic bound for $\eta \leq 1/3$. There are $2n$ calls, each consuming two input tokens and returning one output token, for totals of $4n$ input tokens and $2n$ output tokens. □

B Same-Model False-Positive Rates

Following MET (Gao et al., 2025), we evaluate false positives using same-model comparisons. The false-positive rate (FPR) is the fraction of audits that reject equality when both samples come from the same model. Each comparison of models A and B includes A–A and B–B controls, each evaluated over $R = 25$ repetitions. These controls check whether a method also reports differences on unchanged models.

Table 1 summarizes the same-model controls for the eight model pairs in Figures 3 and 4. Ours and MET use the cheapest plotted configuration reaching 80% empirical power. B3IT uses its plotted highest-power configuration, breaking ties by cost and then k, d, and M. The main-text curves and this summary use power and cost for configuration selection, without screening on empirical FPR. Elevated FPR values are retained in the table.

Table 1: Same-model empirical FPR for the eight model pairs in Figures 3 and 4 at the selected configurations. For each pair A versus B, we report the larger of the A–A and B–B rates, each estimated over $R = 25$ audits. NB denotes no validated B3IT borders.

Model pair	Ours	MET	B3IT
Llama 3.1 8B vs. 70B	0.04	0.16	0.00
Qwen3 8B vs. 30B-A3B	0.00	0.20	NB
Ministral 3 8B vs. 14B	0.04	0.08	0.00
Gemma 3 12B vs. 27B	0.00	0.04	0.76
Llama 3.1 70B vs. Qwen3 30B-A3B	0.00	0.12	0.00
Llama 3.1 70B vs. Ministral 3 14B	0.04	0.12	0.04
Qwen3 30B-A3B vs. Gemma 3 27B	0.00	0.16	0.08
Ministral 3 14B vs. Gemma 3 27B	0.00	0.08	0.00

C Additional Ablation Studies

We use three representative model pairs for all ablations: Llama-3.1-70B-Instruct paired with Llama-3.1-8B-Instruct, Qwen3-30B-A3B-Instruct, or Hermes-3-Llama-3.1-70B. These cover a size change, a cross-family substitution, and a fine-tuned derivative. The panels follow this order in both figures below. We use the same TV statistic, 9,999 paired permutations, nominal significance level $\alpha = 0.05$, and $R = 25$ audits per configuration as in the main experiments. Each curve shows all evaluated query budgets and the corresponding pointwise confidence intervals.

C.1 Impact of Alphabet Size

We vary $K \in \{2, 4, 8, 16, 26\}$ and request one of the first K lowercase English letters. The prompt specifies the range beginning with a, and the parser uses the corresponding alphabet. The response length remains one character, and the statistical test is unchanged. This comparison examines how the number of allowed symbols affects the cost of distinguishing the models.

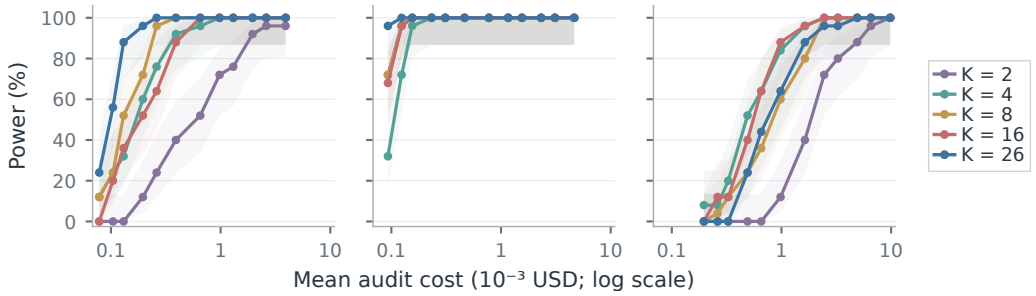


Figure 7: Alphabet-size ablation with $K \in \{2, 4, 8, 16, 26\}$. Each query requests one lowercase letter from the specified alphabet.

Figure 7 shows that the effect of alphabet size varies across model pairs. The full alphabet performs well for the size-change comparison, while intermediate sizes are competitive for the fine-tuned derivative. Restricting the output to two letters requires more expenditure to reach high power in both comparisons. The cross-family pair is easier to distinguish across the evaluated alphabet sizes. These results suggest that reducing the number of possible outcomes can also weaken the model-specific differences retained by the query.

C.2 Impact of Response Length

We fix the alphabet to the 26 lowercase letters and request a random string of $L \in \{1, \ldots, 10\}$ letters, with the output-token cap set to L. After normalization, whitespace stripping, and lowercasing, a valid response must contain exactly L allowed letters. Each complete valid string is treated as one categorical outcome, and all other responses map to INVALID. Cost is computed from realized token usage. Thus, the comparison evaluates each complete response as the fingerprint at the requested length.

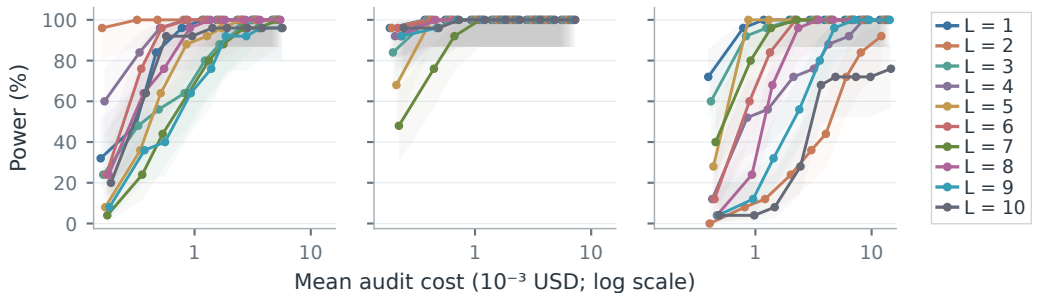


Figure 8: Response-length ablation. Queries request $L = 1, \ldots, 10$ lowercase letters, with an output-token cap of L.

Figure 8 shows no consistent improvement as the requested response length increases. A single character is competitive for the size-change and cross-family comparisons. Longer responses can improve efficiency for the fine-tuned derivative, but the benefit depends on the chosen length. This variation suggests that extra characters are useful when the additional response information offsets their generation cost. The results support one character as a simple, low-cost default, while leaving room for pair-specific benefits from longer responses.