Total Variation Distance Estimation in Autoregressive Models

Eric Price [1] **Kevin Tian** [1] **Zhiyang Xun** [1] **Yusong Zhu** [1]

Abstract

Modern LLM deployments use a number of implementation choices and inference optimizations (e.g., batching, custom kernels, quantization) on top of fixed weights, so two engines serving "the same model" can produce meaningfully different distributions. We study the problem of estimating the total variation (TV) distance between two length-n autoregressive distributions to additive error ε under sample, logit, and noisy-logit access models. Under sample access, we use $\widetilde{O}(n^2K/\varepsilon^2)$ queries, where K is the maximum support of the next-token distribution; under logit access, we use $O(n/\varepsilon^2)$ queries, and this is tight; under noisy logit access, if probability values are given to σ relative error, we use $\widetilde{O}((n+n^2\sigma^2)/\varepsilon^2)$ queries. We complement our theoretical results with an empirical evaluation of our algorithms, for example measuring the distance between `sglang` and `vllm` in standard settings. Our experiments highlight the robustness and practicality of estimating total variation distance compared to alternatives such as KL divergence.

1. Introduction

While the behavior of a large language model (LLM) is often idealized as being determined by its model weights, in practice a wide variety of choices — GPUs, batching, kernels, and low-level numerical effects — can noticeably change its output behavior (He & Thinking Machines Lab, 2025). Other inference-time optimizations, e.g., weight quantization (Frantar et al., 2022; Lin et al., 2024), KV quantization (Liu et al., 2024; Hooper et al., 2024), and lossy speculative decoding methods (Zhou et al., 2024; Li et al., 2024), deliberately introduce approximation errors in exchange for greater efficiency. Other times, it has been documented that model providers simply serve with bugs (Chu et al., 2025). More broadly, different providers of the same open-weight models can exhibit significant differences in latency, throughput, cost, and performance (Artificial Analysis, 2026; Clark, 2025).

These considerations motivate the problem of eliciting differences between two LLMs. We approach this issue from the perspective of *distribution testing*: under an appropriate notion of API access to two distributions μ, π over Σ^n, representing LLMs generating length-n sequences over a token alphabet Σ, can we estimate the distance between μ, π? From a practical perspective, such an estimator could then be used by a user for comparing model providers (how close is the cheap provider to the expensive one?), and conversely, a model provider can use the estimator to test how well a candidate more-efficient inference stack matches their old stack.

In this paper, to evaluate such distributional changes, we focus on measuring the *total variation (TV) distance* between two LLMs. The TV distance, whose definition is recalled in (2), is a convenient and interpretable statistical distance. For example, bounding the TV distance also directly implies bounds on the change to any distinguishing test, e.g., the loss on a fixed evaluation metric.

Prior work on evaluating distributional changes in LLMs (Fireworks AI, 2024; Dutta et al., 2024) often measure the KL divergence, whose definition is also recalled in (2). From a practical standpoint, this has several disadvantages. The KL divergence between distributions is unbounded when their supports are disjoint, so it is not possible to estimate in general to ε precision from finite samples, as this would require support recovery. Smoothed approximations to the KL can be used by e.g., introducing pseudocounts, but these strategies do not have a uniform approximation guarantee and require hyperparameter selection. In contrast, the TV distance is bounded in $[0, 1]$, and therefore it inherits an interpretable scale. Moreover, the TV distance is robust to small implementation differences, such as restricting next-token distributions to their top-k elements for $k = 20$ vs. $k = 30$, or censoring swear words. Such differences alter the support and thus cause KL to diverge. Measuring KL has thus far remained the practitioner default largely because of the chain rule: the sequence-level KL divergence is the sum of the expected per-token KL divergences, so

[1]University of Texas at Austin. Correspondence to: Zhiyang Xun <zxun@cs.utexas.edu>, Yusong Zhu <zhuys@utexas.edu>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

per-step logit comparisons give an unbiased estimator. TV has no such decomposition, which is the technical obstacle this paper addresses.

For further discussion on the impact of the choice of distance metric, see Appendix E.

1.1. Our results

We describe results under three models of accessing LLMs: sample access, logit access, and noisy distribution access. These are formalized in Definitions 1, 2, and 3, but we informally describe them here. Throughout, we fix two distributions π, μ over Σ^n whose TV distance we want to estimate, and we define $m := |\Sigma|$. We also let $K \in [m]$ denote the maximum support size of any next-token distribution $\pi(x_i \mid x_{<i})$ or $\mu(x_i \mid x_{<i})$, where $x_{<i} \in \Sigma^{i-1}$ is a specified prefix. LLM APIs typically support specifying a $k \in \mathbb{N}^+$ and sampling from the top-k tokens; due to nondeterminism in which tokens reach the threshold, the total support size K is somewhat larger than k; but we find it to be well below $2k$ in practice (Figure 6 in Appendix D.5).

Sample access. The simplest and most general access model we consider is *sample access*: we can query any prefix $x_{<i}$ and get a sample from the next-token distribution $\pi(x_i \mid x_{<i})$. This oracle is explicitly studied by (Adar et al., 2024; Meel et al., 2025), and has an intellectual history in the recent theoretical computer science literature on conditional property testing (Chakraborty et al., 2013; Acharya et al., 2014; Canonne et al., 2015; Bhattacharyya & Chakraborty, 2018).

Theorem 1 (Informal, cf. Corollary 1)**.** *There is an algorithm that estimates* $\mathrm{TV}(\pi, \mu)$ *up to error* ε *with probability* $\geq \frac{9}{10}$*, using* $O(\frac{n^2 K}{\varepsilon^2} \cdot \log^2(\frac{1}{\varepsilon}))$ *queries to prefix sampling oracles to both* π *and* μ*.*

The most directly-related result to Theorem 1 is by (Meel et al., 2025), whose Theorem 1 uses $O(\frac{n^3 m}{\varepsilon^5})$ prefix sampling oracle queries. Theorem 1 is better in terms of sequence length n, accuracy ε, and it saves an $\frac{m}{K}$ factor when the support is sparse[1] . This $\frac{m}{K}$ factor can be quite large: for reference, Qwen has m=151,643 tokens and recommends sampling with $k = 20$ (Qwen Team, 2026); such parameters are typical.

Logit queries. Beyond providing sample access, many LLM APIs also return the logits: we can query any prefix $x_{<i}$ and see the full vector $\pi(x_i \mid x_{<i})$. This is much more powerful—for instance, it lets us compute $\pi(x)$ with n queries—leading to a better sample complexity.

Theorem 2 (Informal, cf. Theorems 4, 5)**.** *There is an algorithm that estimates* $\mathrm{TV}(\pi, \mu)$ *up to error* ε *with probability* $\geq \frac{9}{10}$*, using* $O(\frac{n}{\varepsilon^2})$ *queries to prefix logit oracles to both* π *and* μ*. Moreover, this oracle query complexity is tight up to constant factors for any algorithm.*

To our knowledge we are the first paper to formalize the corresponding TV distance estimation problem under logit access. As discussed, the logit variant of the problem is natural in practical settings where the entire next-token sampling distribution can be queried, and Theorem 5 provides matching upper and lower bounds for its query complexity.

Our upper bound is proven using a simple observation: $O(n)$ queries to prefix logits allows exact computation of a likelihood-ratio $\frac{\pi(x)}{\mu(x)}$, for any sequence $x \in \Sigma^n$. This allows us to repurpose a simple TV estimator from (Canonne & Rubinfeld, 2014), which uses $O(\frac{1}{\varepsilon^2})$ likelihood-ratio queries. We complement this algorithm with a matching distribution over hard instances π and μ, indexed by a hidden parameter $p \in [0, 1]$. For this construction, we reduce estimating $\mathrm{TV}(\pi, \mu)$ to learning p up to additive error ε. By carefully defining the hard instance, we show that $\approx n$ prefix logit oracle queries are needed to simulate one toss $\sim \mathrm{Bern}(p)$, at which point $\approx \frac{1}{\varepsilon^2}$ tosses are necessary.

Noisy logits. Unfortunately, while LLM APIs do return logits, those logits are often not reliable. The issues (batching, kernels, precision) that motivate this paper by making LLM implementations unreliable can also make the computation of logits non-deterministic. In Figure 1 we see that some popular open-source LLM implementations return probabilities that regularly vary by 5–10%.

To model this, for any given prefix $x_{<i}$ we can no longer query the vector $p := \pi(x_i \mid x_{<i})$ directly. Instead, we suppose that we can sample from an unbiased estimator with bounded relative variance; that is, we can sample a random vector $\overline{p}$ such that $\mathbb{E}[\overline{p}] = p$ and

$$\mathbb{E}_{a \sim p}\left[\mathbb{V}\mathrm{ar}\left[\frac{\overline{p}_a}{p_a}\right]\right] \leq \sigma^2. \tag{1}$$

This is implied, for instance, by estimating the final logits (before the softmax) to additive $\Theta(\sigma^2)$-variance Gaussian noise, see Lemma 5; and it is empirically validated in Figure 1 (a).

We believe this to be a fairly reasonable model. The challenge and non-determinism in sampling from an LLM comes from computing the logits, not in taking the final softmax; whatever process produces the logits and corresponding $\overline{p}$, the average of $\overline{p}$ really will be the true p. The main caveat is that we need independence across samples.

This "noisy prefix distribution model" *smoothly interpolates*

[1] Depending on m rather than K appears inherent to the (Meel et al., 2025) strategy, which involved "taming" distributions by padding with a $\frac{1}{m}$-scaled uniform distribution to ensure their estimator converges (see discussion after their Theorem 4.8).

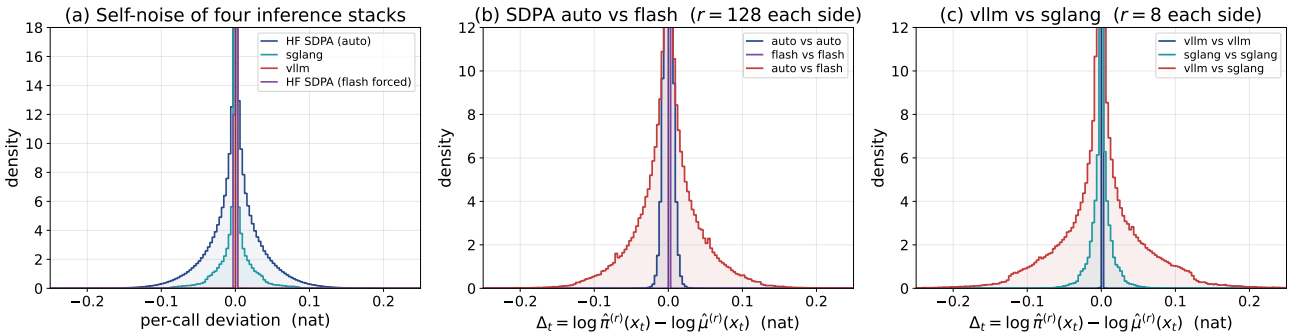

Figure 1. (a) The logits returned by popular inference engines are often not deterministic. For example, the probabilities `sglang` returns typically vary by 5%; but PyTorch SDPA with the Flash Attention 2 backend is deterministic. (b,c) After averaging over several draws, we converge toward the "true" probability; different inference engines converge to systematically different probabilities (e.g. sglang and vllm are typically 5-10% different).

between the previous models (cf. Lemma 4). If $\sigma = 0$, it is precisely the logit query model. With just sample access, we can simply consider the sample $a \sim p$ as the one-hot distribution e_a. This is unbiased with variance (1)

$$\begin{aligned}\sigma^2 &:= \sum_{a \in \mathrm{supp}(p)} \frac{\mathrm{Var}[\overline{p}_a]}{p_a} \\ &= \sum_{a \in \mathrm{supp}(p)} (1 - p_a) \\ &= |\mathrm{supp}(p)| - 1 < K.\end{aligned}$$

Our last main result gives an efficient query complexity bound under noisy logit access.

Theorem 3. *There is an algorithm that estimates* $\mathrm{TV}(\pi, \mu)$ *up to error* ε *with probability* $\geq \frac{9}{10}$*, using* $O(\frac{n+n^2\sigma^2}{\varepsilon^2} \log^2(\frac{1}{\varepsilon}))$ *queries to noisy prefix distribution oracles for both* π *and* μ*.*

Theorem 3 precisely recovers Theorem 1, and it recovers Theorem 2 up to $\log^2 \frac{1}{\varepsilon}$.

To prove Theorem 3, we first consider an alternative representation of the TV distance (cf. (6) in Lemma 1), which allows us to avoid the "distribution taming" step caused by the unbounded estimators used by the (Meel et al., 2025) algorithm. As we discuss in Section C.2, this simple change already improves the query complexity of (Meel et al., 2025) to $\approx \frac{n^2 K}{\varepsilon^4}$. To achieve our tightest result in Section C.3, we use a variance reduction scheme patterned off of multilevel Monte Carlo (MLMC) estimation, a classical technique in statistics (Heinrich, 2001; Giles, 2015). Our approach is particularly inspired by an MLMC scheme from (Giles, 2008), which trades off cheaper query complexities with larger variances at different levels of estimation, using a telescoping sum.

We summarize our main theoretical results in Table 1, which illustrates the tradeoff between Theorems 1 and 2 that is interpolated in Theorem 3: the stronger logit query model improves the query complexity by a factor of n and eliminates the dependence on K. An intriguing question which we leave open for future work is whether this separation is inherent, i.e., if superlinear-in-n lower bounds or subquadratic-in-n upper bounds can be proven for the prefix sampling oracle access model.

Experiments. We empirically evaluate Algorithm 3 in Section 4. On a synthetic autoregressive instance with known true TV, the algorithm returns the correct value across a range of noise levels, and a measurable diagnostic guides the choice of multilevel depth on each instance. We then apply it to a real measurement of independent interest: the TV between two production inference engines, `vllm` and `sglang`, serving identical `Qwen3-0.6B` weights; the result and a finer decomposition are summarized in Figure 2.

1.2. Related work

TV estimation under conditional access. TV estimation is a quantitative analogue of distribution testing: an additive estimator immediately yields tolerant and non-tolerant testers by thresholding. With i.i.d. samples, even non-tolerant identity testing over $\{0,1\}^n$ requires $\Omega(2^{n/2})$ samples (Chan et al., 2014; Valiant & Valiant, 2017), which motivates stronger sampling interfaces. Conditional sampling oracles have been extensively studied in theoretical computer science (TCS), from full conditioning on arbitrary events (Chakraborty et al., 2013) to more restricted oracles such as pair (Acharya et al., 2014), interval (Canonne et al., 2015; Narayanan, 2020), subcube (Bhattacharyya & Chakraborty, 2018; Canonne et al., 2021; Chakrabarty et al., 2025), and prefix (Bhattacharyya et al., 2024; Adar et al., 2024) oracles. Subcube and prefix models are closest to the autoregressive sampling setting. However, most prior work focuses on non-tolerant testing rather than distance estimation; see (Canonne, 2020) for a broader survey.

Oracle	Prior upper	Our upper	Prior lower	Our lower
Sampling	$\frac{n^3 m}{\varepsilon^5}$ (Meel et al., 2025)	$\frac{n^2 K}{\varepsilon^2}$	$\frac{n}{\log n}$ (Canonne et al., 2020)	$\frac{n}{\varepsilon^2}$
Logit	–	$\frac{n}{\varepsilon^2}$	–	$\frac{n}{\varepsilon^2}$
Noisy logit	–	$\frac{n + n^2\sigma^2}{\varepsilon^2}$	$\frac{n}{\log n}$ (Canonne et al., 2020)	$\frac{n}{\varepsilon^2}$

Table 1. Prefix query complexity bounds for TV estimation over Σ^n with logarithmic factors omitted. Here $m := |\Sigma|$, and $K \leq m$ is the maximum support size of any next-token distribution.

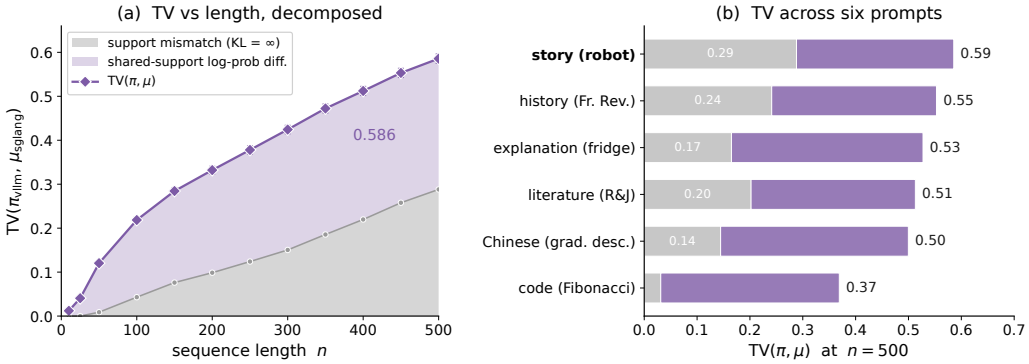

Figure 2. Two production inference engines, `vllm` and `sglang`, serving identical `Qwen3-0.6B` weights (bf16, top-$k = 20$) are far apart at face value: $\mathrm{TV}(\pi, \mu) = 0.586 \pm 0.011$ at sequence length $n = 500$ on the reference prompt. Scoring uses a teacher-forced *replay oracle* that re-decodes each sampled trajectory (forcing each step to the pre-sampled token), reproducing the engine's own sampling distribution without bias. *(a)* The estimate decomposes into a *shared-support log-probability difference* (both engines keep the sampled token) and a *support mismatch* (one engine assigns zero r-averaged top-k mass to a token the other sampled, where an empirical KL would return $+\infty$); the mismatch term grows with n and accounts for 49% of the distance at $n = 500$. *(b)* The measurement is not specific to one prompt: across six prompts spanning story, explanation, history, literature, code, and Chinese content, $\mathrm{TV}(\pi, \mu) \in [0.37, 0.59]$, ordered by the content's token entropy (code lowest).

The closest theoretical works to ours study TV estimation under prefix sampler access. On the upper-bound side, (Meel et al., 2025; Bhattacharyya et al., 2024; Adar, 2025) consider essentially the same sampler oracle as ours. (Meel et al., 2025) show that $O(n^3 m/\varepsilon^5)$ queries suffice, while (Bhattacharyya et al., 2024) improve the bound to $O(n^2 m/\varepsilon^4)$ when one of π, μ is known. Using a different approach, (Adar, 2025) obtain an $O(n^2/\varepsilon^4)$ bound, but only for $m = 2$. On the lower-bound side, (Canonne et al., 2020) prove an $\Omega(n/\log n)$ lower bound for non-tolerant testing, which also implies hardness for TV estimation. In addition, (Adar, 2025) prove an $\Omega(n^2/\varepsilon^2)$ lower bound for elementwise probability mass estimation. Although this does not rule out an $o(n^2)$ TV estimator, it suggests that sharper bounds may require new techniques. A separate line of work studies TV estimation with numerical probability access: (Canonne & Rubinfeld, 2014; Bhattacharyya et al., 2020) assume direct queries to $P(x)$ and $Q(x)$ and show that $\Theta(1/\varepsilon^2)$ queries are necessary and sufficient. Their algorithms can be adapted to our prefix-logit setting, yielding an $O(n/\varepsilon^2)$-query algorithm, but their lower bounds do not apply because their hard instances do not capture prefix structures.

Auditing distribution shifts in LLMs. Auditing whether an LLM API serves the model it claims to provide has become a central question in trustworthy model deployment. Some works approach this problem through downstream task performance (Eyuboglu et al., 2024; Chen et al., 2024). However, task accuracy is only an indirect proxy for model substitution: two models may achieve similar correctness while inducing substantially different output distributions, leading to changes in token usage, response style, or generation quality. A more direct approach is therefore to audit the model-induced output distribution itself, where the core question becomes whether, and by how much, the served distribution differs from the claimed one. Several recent works take this distributional perspective. (Gao et al., 2024) proposes a model-equality test, corresponding to non-tolerant testing in TCS, and uses MMD as the test statistic. While empirically promising, this approach lacks theoretical guarantees and was later found to be less effective under subtle substitutions, such as lower-precision inference or partial-time model replacement (Cai et al., 2025). (Zhu et al., 2025)

strengthens detection via rank-based nonparametric tests, but its complexity and significance analyses rely on assumptions on the log-ranks of model outputs. Finally, (Amini et al., 2025) proposes an algorithm for estimating KL divergence between LLMs, but its analysis assumes bounded estimator variance, which can be very large or even infinite.

2. Preliminaries

In this section, we introduce preliminaries used throughout the paper. Section 2.1 fixes notation; and Section 2.2 defines the oracles used to describe autoregressive models.

2.1. Notation

Probability. For a state space Ω, we let $\mathcal{P}(\Omega)$ denote all probability measures over Ω. For an event $\mathcal{E} \subseteq \Omega$, we let $\mathbb{I}(\mathcal{E})$ denote the corresponding 0-1 indicator random variable, and $\mu(\mathcal{E})$ denote the probability of the event. We let $\mathbb{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ denote the expectation and variance. We frequently use the following distances (and divergences) between two distributions $\mu, \pi \in \mathcal{P}(\Omega)$, where $\mathrm{d}\omega$ denotes a counting measure if Ω is discrete:

$$\begin{aligned} \mathrm{TV}\left(\mu, \pi\right) &:= \frac{1}{2}\int_{\Omega} \left|\mu(\omega) - \pi(\omega)\right| \mathrm{d}\omega \\ &= \sup_{\mathcal{E}\subseteq\Omega} \mu(\mathcal{E}) - \pi(\mathcal{E}), \\ \mathrm{KL}\left(\mu \| \pi\right) &:= \int_{\Omega} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega)\right)\mu(\omega)\mathrm{d}\omega, \\ \chi^2\left(\mu \| \pi\right) &:= \int_{\Omega} \left(\frac{\mu(\omega)}{\pi(\omega)} - 1\right)^2 \pi(\omega)\mathrm{d}\omega. \end{aligned} \tag{2}$$

Autoregressive models. Throughout, we let Σ be a finite alphabet with $m := |\Sigma|$, and let Σ^n denote the space of length-n sequences. In an LLM setting, Σ represents a tokenizer vocabulary, and n is a sequence length. When comparing two distributions π and μ, we reserve use of the letter $K \in [m]$ to denote the maximum support size of any (prefix-conditional) next-token distribution, i.e., $\pi(X_{i+1} = \cdot \mid X_1, \ldots, X_i)$ or $\mu(X_{i+1} = \cdot \mid X_1, \ldots, X_i)$. Next-token distributions in $\mathcal{P}(\Sigma)$ are typically implicitly specified as $\mathrm{softmax}(\ell)$ for logits $\ell \in \mathbb{R}^m$, where $\mathrm{softmax}(\ell) \propto \exp(\ell)$.

Finally we introduce some notation for manipulating sequences. For $x \in \Sigma^n$ and $0 \le i \le n$, we write $x_{1:i} = (x_1, \ldots, x_i)$, with $x_{1:0} = \emptyset$. For $\pi \in \mathcal{P}(\Sigma^n)$, let $\pi_{1:i}$ denote the marginal distribution of $X_{1:i}$. We also use $\pi_{x_{1:i}}(a)$ as shorthand for the next-token distribution $\pi(X_{i+1} = a \mid X_{1:i} = x_{1:i})$. For sequences $s \in \Sigma^m$ and $s' \in \Sigma^n$, let $s \oplus s' \in \Sigma^{m+n}$ denote their concatenation. When the domain is clear, we use e_a to denote the ath basis vector; typically, this is used when $a \in \Sigma$ so that $e_a \in \mathcal{P}(\Sigma)$.

2.2. Oracle definitions

We introduce three oracle models for autoregressive distributions, corresponding to different levels of access to an autoregressive model. The strongest model is exact prefix logit access. This corresponds to a white-box setting: one can run the model forward pass at a given prefix and read off the next-token log-probabilities, as if the model weights and inference procedure were directly available.

Definition 1 (Prefix logit oracle)**.** *Let $\pi \in \mathcal{P}(\Sigma^n)$. For any prefix $x_{1:i} \in \Sigma^i$ with $0 \le i < n$, the prefix logit oracle returns the vector of next-token log-probabilities* [2]

$$\mathcal{O}^{\pi}_{\mathrm{logit}}(x_{1:i}) := \left(\log \pi(X_{i+1} = a \mid X_{1:i} = x_{1:i})\right)_{a\in\Sigma}. \tag{3}$$

We also use the notation

$$\mathcal{O}^{\pi}_{\mathrm{logit}}(x_{1:i}, a) := \log \pi(X_{i+1} = a \mid X_{1:i} = x_{1:i}), \qquad a \in \Sigma. \tag{4}$$

It is clear that one query to the vector oracle (3) simulates one scalar oracle query (4) at no additional cost. In the reverse direction, a scalar oracle simulates the full vector oracle using m scalar queries. We ignore this alphabet-size distinction in our main complexity statements under prefix logit oracle access. This convention is natural for LLMs, where a single forward pass produces all logits simultaneously, and log-probabilities differ from logits only by a softmax normalization.

The second model is prefix sampling access. This model is a special case of the more general conditional sampling oracle model that has received substantial attention in the theoretical computer science literature (Chakraborty et al., 2013; Acharya et al., 2014; Canonne et al., 2015; Bhattacharyya & Chakraborty, 2018). Under this model, the algorithm cannot directly inspect the model logits, but it can request a token sampled from the model's next-token conditional distribution. This is also a natural abstraction for closed-source or commercial language models that do not allow explicit access under Definition 1.

Definition 2 (Prefix sampler oracle)**.** *Let $\pi \in \mathcal{P}(\Sigma^n)$. For any prefix $x_{1:i} \in \Sigma^i$ with $0 \le i < n$, the prefix sampler oracle returns*

$$\mathcal{O}^{\pi}_{\mathrm{samp}}(x_{1:i}) \sim \pi(\cdot \mid X_{1:i} = x_{1:i}).$$

Finally, we introduce a noisy prefix logit oracle in Definition 3. This is the most general oracle considered in this paper, subsuming both exact prefix logit queries and prefix sampling queries as special cases. The oracle returns

[2] If $\pi(X_{1:i} = x_{1:i}) = 0$, the conditional distribution is undefined; in this case, one may define the oracle to return a distinguished symbol $\perp$. Throughout the paper, unless explicitly specified, our analysis will not invoke it.

an unbiased random sketch of the next-token distribution, with a noise level parameterized by relative variance. Beyond being a convenient parameterization for analysis, this model captures self-inconsistency in LLM inference, where inference-time randomness can perturb the logits and induce different next-token distributions for the same prefix. We substantiate these interpretations in Appendix C.1, and highlight the practical relevance of modeling query noise in Section 4, when we test the distance between an autoregressive model and itself.

Definition 3 (Noisy prefix distribution oracle)**.** *Let $\pi \in \mathcal{P}(\Sigma^n)$. For any prefix $x_{1:i} \in \Sigma^i$ with $0 \le i < n$, the noisy prefix distribution oracle with relative variance σ^2 returns an unbiased vector $\mathcal{O}^{\pi}_{\mathrm{ndist}}(x_{1:i}) \in \mathcal{P}(\Sigma)$. Writing $p_i = \pi_{x_{1:i}}$ and $\widehat{p} = \mathcal{O}^{\pi}_{\mathrm{ndist}}(x_{1:i})$, the oracle satisfies $\mathbb{E}[\widehat{p}] = p_i$ and*

$$\mathbb{E}\left[\chi^2\left(\widehat{p}\|p_i\right)\right] = \mathbb{E}_{a\sim p_i}\left[\mathrm{Var}\left(\frac{\widehat{p}(a)}{p_i(a)}\right)\right] \le \sigma^2,$$

where $\mathbb{E}$ on the left and Var *on the right are over the internal randomness of $\mathcal{O}_{\mathrm{ndist}}$.*

Lemma 4 proves that Definition 2 is a case of Definition 3 with $\sigma^2 \le K$. Notably, Definition 2 is the hardest case of Definition 3, so we may always assume that $\sigma^2 \in [0, K]$. This is because we can always simulate Definition 2 by querying $a \sim \mathcal{O}_{\mathrm{ndist}}(x_{1:i})$ and outputting e_a, because $\mathbb{E}[\mathcal{O}_{\mathrm{ndist}}(x_{1:i})] = \pi_{x_{1:i}}$. We also give broader motivation for our relative variance parameterization in Lemma 5.

We use the following unified formulation for our main distance estimation problem under all three oracle models. The distributions π and μ are only accessed through their corresponding oracles, and the goal is to estimate their TV distance to additive accuracy ε.

Problem 1 (TV estimation)**.** *Let $\pi, \mu \in \mathcal{P}(\Sigma^n)$. Assume access to appropriate prefix oracles for μ and π, denoted $\mathcal{O}^{\pi}, \mathcal{O}^{\mu}$. Given $\varepsilon, \delta \in (0, 1)$, design an algorithm that outputs $\widehat{d}$ such that*

$$\mathbb{P}\left[\left|\widehat{d} - \mathrm{TV}\left(\mu, \pi\right)\right| \le \varepsilon\right] \ge 1 - \delta.$$

The query complexity is the total number of calls to $\mathcal{O}^{\pi}$ and $\mathcal{O}^{\mu}$.

Across our analysis, we instantiate $\mathcal{O}^{\pi}$ and $\mathcal{O}^{\mu}$ with the access models in Definitions 1, 2, and 3.

3. Technical Overview

In this section we give an overview of our main theoretical results, particularly our upper bound Theorem 6 and our lower bound Theorem 5. For associated proofs, we refer to Appendix B and C.

3.1. Estimator construction

We first describe our estimator under noisy prefix distribution access. The starting point is a mixture representation of TV distance that we recall in (6):

$$\begin{aligned} \mathrm{TV}\left(\pi, \mu\right) &= \mathbb{E}_{X\sim M}[Z(X)], \\ M &= \frac{\pi + \mu}{2}, \\ Z(x) &= \frac{|\pi(x) - \mu(x)|}{\pi(x) + \mu(x)}, \quad x \in \mathrm{supp}(\pi) \cup \mathrm{supp}(\mu). \end{aligned}$$

Thus, it suffices to estimate the local discrepancy $Z(X)$ for strings $X \sim M$. Sampling from M is straightforward: we first choose one of π, μ uniformly at random, and then roll out a string token by token using n queries to the corresponding prefix oracle.

For a fixed string x, the probability $\pi(x)$ factors into the product of its prefix conditionals. We estimate each conditional probability by averaging r independent noisy oracle outputs at the corresponding prefix, and then multiply these averaged conditionals to obtain a plug-in estimate $\widehat{\pi}_r(x)$. The estimate $\widehat{\mu}_r(x)$ is defined analogously. We then estimate $Z(x)$ by replacing $\pi(x)$ and $\mu(x)$ with $\widehat{\pi}_r(x)$ and $\widehat{\mu}_r(x)$ in the normalized discrepancy formula.

The main stability issue is whether this product estimator accumulates noise too quickly across the n prefixes. We control this through the χ^2 error, which captures the relative squared error of the trajectory probability estimate. For a trajectory x, the ratio $\widehat{\pi}_r(x)/\pi(x)$ is a product of n one-step relative estimates. Each one-step estimate has second moment $1 + O(\sigma^2/r)$, and independence across prefixes gives

$$\mathbb{E}\big[\chi^2(\widehat{\pi}_r\|\pi)\big] \lesssim (1 + O(\sigma^2/r))^n - 1 = O(n\sigma^2/r),$$

provided $r \gtrsim n\sigma^2$. Hence the product structure causes only a linear accumulation of relative squared error, rather than an exponential one.

This relative-error control is also well matched to the normalized discrepancy $Z(x)$. Although the denominator $\pi(x) + \mu(x)$ can be small, the mixture distribution $M = (\pi + \mu)/2$ downweights such strings by the same quantity. Consequently, the mean-squared error of the plug-in discrepancy is $O(n\sigma^2/r)$. Taking $r = O(n\sigma^2/\varepsilon^2)$ makes the plug-in approximation error $O(\varepsilon)$. A direct Monte Carlo estimator of $\mathbb{E}[\widehat{Z}_r(X)]$ would then require $O(1/\varepsilon^2)$ outer samples, giving suboptimal query complexity $O(n^2\sigma^2/\varepsilon^4)$, a factor $\approx \varepsilon^{-2}$ away from our final bound.

3.2. Variance reduction

We use a multilevel decomposition because directly estimating the finest-level approximation $\theta_L = \mathbb{E}[\widehat{Z}_L(X)]$ is

wasteful: each sample of $\widehat{Z}_L$ is expensive, since it uses many oracle repetitions at every prefix. Instead, we decompose θ_L into a cheap coarse estimate plus a sequence of refinement terms:

$$\begin{aligned}\theta_L &= \theta_0 + \sum_{\ell=1}^{L}(\theta_\ell - \theta_{\ell-1}) \\ &= \mathbb{E}[\widehat{Z}_0(X)] \\ &\quad + \sum_{\ell=1}^{L}\mathbb{E}\Big[\widehat{Z}_\ell(X) - \widehat{Z}_{\ell-1}(X)\Big].\end{aligned}$$

Here $\theta_\ell = \mathbb{E}[\widehat{Z}_\ell(X)]$, and larger ℓ means a more accurate but more expensive plug-in estimator. Thus, each difference $\theta_\ell - \theta_{\ell-1}$ only measures the additional improvement gained by increasing the number of oracle repetitions from level $\ell - 1$ to level ℓ. The key point is that each refinement term has much smaller variance than $\widehat{Z}_\ell(X)$ itself (which is dominated by the intrinsic variance of Z). To see this, we estimate $\theta_\ell - \theta_{\ell-1}$ by drawing one shared sample $X \sim M$ and computing

$$Y_\ell = \widehat{Z}_\ell(X) - \widehat{Z}_{\ell-1}(X).$$

Because both terms are evaluated on the same X, the underlying signal $Z(X)$ cancels:

$$Y_\ell = \big(\widehat{Z}_\ell(X) - Z(X)\big) - \big(\widehat{Z}_{\ell-1}(X) - Z(X)\big).$$

Therefore, Y_ℓ is governed only by the approximation errors of the two adjacent levels. As the level increases, these errors shrink, and hence the variance of the coupled increment decays geometrically. This is where the telescoping scheme saves queries: coarse corrections are cheap but noisy, while fine corrections are expensive but have much smaller variance, so we can allocate samples nonuniformly across levels. Concretely, with $r_\ell = n\sigma^2 2^\ell$, one sample from level ℓ costs $O(nr_\ell) = O(n^2\sigma^2 2^\ell)$, whereas the coupled variance satisfies $\mathrm{Var}(Y_\ell) = O(2^{-\ell})$. Thus it suffices to take $O(2^{-\ell}/\varepsilon^2)$ samples at level ℓ. The total cost of this level is therefore

$$O(n^2\sigma^2 2^\ell)\cdot O\left(\frac{2^{-\ell}}{\varepsilon^2}\right) = O\left(\frac{n^2\sigma^2}{\varepsilon^2}\right).$$

The increasing per-sample cost and decreasing variance exactly offset each other, so each level contributes roughly the same amount. Summing over $L = O(\log(1/\varepsilon))$ levels gives the final query bound up to logarithmic factors, which are incurred due to the levels of the scheme.

3.3. Lower bound

We prove the lower bound in Theorem 5 by embedding a Bernoulli parameter-estimation problem into TV estimation. For simplicity, consider the binary alphabet. We split a length-n string as $x = u \oplus z$, where the first r bits u identify one of $L = 2^r$ blocks and the remaining $t = n - r$ bits specify a leaf inside that block. In each block u, we draw a hidden bit $X_u \sim \mathrm{Bern}(p)$. If $X_u = 0$, the two distributions π and μ are identical inside the block; if $X_u = 1$, they are maximally separated, but only at a hidden pair of sibling leaves. Thus the global TV distance decomposes as $\mathrm{TV}(\pi,\mu) = \frac{1}{L}\sum_{u\in\{0,1\}^r} X_u$. For $L \gtrsim 1/\varepsilon^2$, this average is concentrated around p. Therefore, an ε-accurate TV estimator can distinguish $p = p_0$ from $p = p_1$, where $p_1 - p_0 = \Theta(\varepsilon)$.

The construction hides each bit X_u behind a random suffix prefix. More concretely, within block u, the two distributions only reveal whether they are identical or separated when the algorithm queries the unique hidden parent of the sibling leaves. Before reaching this parent, oracle answers only provide limited information about where the hidden parent is; they do not reveal X_u. Hence learning one useful Bernoulli sample X_u requires locating a hidden string of length $t - 1$ inside the block. The appendix formalizes this through a transcript mutual-information argument: any adaptive algorithm making B prefix-logit queries can reveal only $O(B/t)$ such blocks in expectation.

Now suppose a TV estimator used only B oracle queries. We could simulate its interaction with the hard instance using an unknown coin of bias $p \in \{p_0, p_1\}$, lazily flipping the coin only when the transcript actually reveals some X_u. Since the transcript reveals only $O(B/t)$ blocks, this would give a distinguisher between $\mathrm{Bern}(p_0)$ and $\mathrm{Bern}(p_1)$ using only $O(B/t)$ coin samples. But distinguishing two Bernoulli biases separated by $\Theta(\varepsilon)$ requires $\Omega(1/\varepsilon^2)$ samples. Therefore $B = \Omega(t/\varepsilon^2)$. Taking $r = O(\log(1/\varepsilon))$ ensures concentration of the block average, and when n is sufficiently larger than r, we have $t = \Theta(n)$. This yields the lower bound $\Omega(n/\varepsilon^2)$.

4. Experiments

We empirically evaluate Algorithm 3 in two settings: a synthetic autoregressive instance with controllable noise, where the true TV is known by direct enumeration (Section 4.1), and a real LLM deployment, where the TV between two production inference engines serving identical model weights is itself a question of independent interest (Section 4.2). Implementation details, instance parameters, and the methodology behind individual quantitative claims are deferred to Appendix D.

4.1. Validation on a synthetic hard instance

We instantiate the binary autoregressive hard instance of Appendix B.2 with controllable calibrated probability noise of standard deviation $\sigma\sqrt{\hat{p}(1-\hat{p})}$ (deterministic transitions

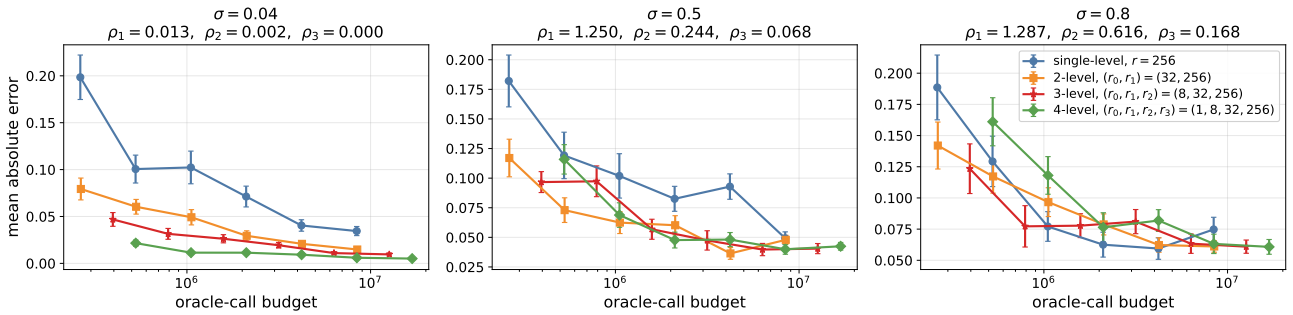

Figure 3. Mean absolute error vs. oracle-call budget for 1-, 2-, 3-, 4-level instantiations of Algorithm 3 on the synthetic hard instance, at three noise levels $\sigma \in \{0.04, 0.5, 0.8\}$. Column titles report the variance ratio ρ_ℓ at each level ℓ of the deepest (4-level) schedule.

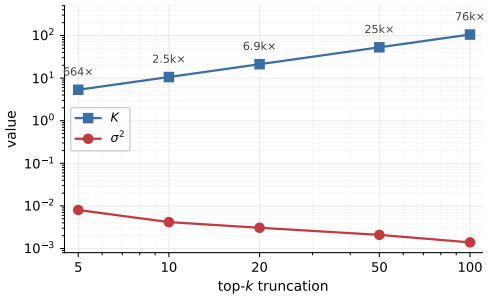

Figure 4. Empirical noisy-oracle parameter σ^2 (Definition 3) and effective support size K on `Qwen3-0.6B` (bf16, sdpa), against the API top-k truncation; annotations show K/σ^2, from $660\times$ at $k = 5$ to 7.6×10^4 at $k = 100$. Methodology in Appendix D.6.

stay deterministic, and the K-average is unbiased for the true probability). The true $\mathrm{TV}(\pi, \mu) = 0.388$ is known by direct enumeration, which lets us ask two questions: *(i)* does Algorithm 3 return the correct value across the noise range, and *(ii)* does its multilevel structure produce instance-wise cost savings beyond the worst-case bound of Theorem 3? All four schedules in Figure 3 converge to the true TV at every σ, settling *(i)*. Question *(ii)* reduces to choosing the right multilevel depth, which we can read off the data through a measurable diagnostic.

The cost benefit of including each level is captured by the empirical variance ratio

$$\rho_\ell := \frac{\mathrm{Var}[Y_\ell]}{\mathrm{Var}[Z_\ell]} = \frac{\mathrm{Var}(Z_\ell - Z_{\ell-1})}{\mathrm{Var}(Z_\ell)},$$

where Y_ℓ and Z_ℓ are the coupled level-ℓ increment and estimator from described formally from (12)–(16). When $\rho_\ell \ll 1$, the cheaper level $\ell - 1$ provides an essentially free variance reduction at level ℓ, and including it strictly beats the single-level estimator at r_ℓ at matched RMSE; when ρ_ℓ is order one, level $\ell - 1$ adds cost without reducing variance and should be dropped.

Figure 3 sweeps $\sigma \in \{0.04, 0.5, 0.8\}$ across four schedule depths under matched query budget, with ρ_ℓ for each level reported in the column titles. The empirical winner tracks ρ_ℓ at every σ: at $\sigma = 0.04$ all $\rho_\ell < 0.02$ and the deepest (4-level) schedule beats the single-level estimator at $r = 256$ by $\approx 7\times$ at matched RMSE; as σ grows, ρ_1 rises past 1 and the optimal depth shrinks; at $\sigma = 0.8$ the 2-level schedule wins. Since ρ_ℓ is itself measurable on a small calibration sample, a practitioner can pick the depth before paying the full estimation budget.

4.2. Application: cross-engine TV between vllm and sglang

We apply Algorithm 3 to a measurement of independent interest: for a single fixed prompt, the TV between the length-n continuation distributions of two production inference engines, `vllm` and `sglang`, serving identical `Qwen3-0.6B` weights (top-$k = 20$, bf16). The two engines differ in attention kernel, batching, and scheduler, but should in principle realize the same next-token distribution at every prefix; the empirical question is how different they are at face value.

Black-box measurement requires a scoring oracle faithful to the engine's own sampling path. A whole-sequence log-probability API is biased: its tie-breaks and execution path disagree with generation, assigning zero mass to a token the engine itself sampled on 37% of `sglang` trajectories (Appendix D). We instead use a teacher-forced *replay oracle*: the engine re-decodes the sampled trajectory under the sampling-time batch shape, forcing each step to the pre-sampled token, and we capture the pre-forcing top-k log-softmax; r replays are then r i.i.d. draws of the noisy oracle.

The estimator returns $\mathrm{TV}(\pi, \mu) = 0.586 \pm 0.011$ at $n = 500$ on the reference prompt (Figure 2(a)). This is exactly the regime where the choice of TV over KL matters: at $n = 500$, on 29% of trajectories one engine assigns zero r-averaged top-k mass to a token the other sampled, so the empirical KL is $+\infty$ there and cannot distinguish a small disagreement from a large one, while Algorithm 3 replaces each such event with the bounded contribution

1 from (12) and returns a finite estimate. This support-mismatch term accounts for 49% of the distance at $n = 500$ and grows with length (to $\approx 48\%$ of trajectories by $n = 2000$). The measurement is not specific to one prompt: across six prompts (Figure 2(b)) $\mathrm{TV}(\pi, \mu) \in [0.37, 0.59]$, ordered by the content's token entropy.

Each engine is independently a noisy oracle, and we calibrate the replay oracle's noise for each. `sglang`'s forced decode is bit-identical across repetitions, so we measure $\sigma_{\text{sglang}} = 0$; `vllm`'s carries a small per-token noise $\sigma_{\text{vllm}} \approx 8 \times 10^{-4}$ (on $\approx 0.1\%$ of tokens). The single-repetition self-noise floor is correspondingly 0 for `sglang` and 0.007 for `vllm` at $n = 500$, both far below the cross-engine $\mathrm{TV}(\pi, \mu) = 0.586$, so a single measurement resolves the distance.

On the same model, σ^2 is empirically 10^3–10^5 times smaller than K shown in Figure 4, so Theorem 3's $(n + n^2\sigma^2)/\varepsilon^2$ bound is far tighter than the sample bound n^2K/ε^2 here.

References

Acharya, J., Canonne, C. L., and Kamath, G. A chasm between identity and equivalence testing with conditional queries. *arXiv preprint arXiv:1411.7346*, 2014.

Adar, T. Tight simulation of a distribution using conditional samples. *arXiv preprint arXiv:2506.18444*, 2025.

Adar, T., Fischer, E., and Levi, A. Improved bounds for high-dimensional equivalence and product testing using subcube queries. *arXiv preprint arXiv:2408.02347*, 2024.

Amini, A., Vieira, T., and Cotterell, R. Better estimation of the kullback–leibler divergence between language models. *arXiv preprint arXiv:2504.10637*, 2025.

Artificial Analysis. gpt-oss-120B (high): API provider performance benchmarking & price analysis. https://artificialanalysis.ai/models/gpt-oss-120b/providers#evaluations, 2026. Accessed: 2026-05-04.

Bhattacharyya, A., Gayen, S., Meel, K. S., and Vinodchandran, N. Efficient distance approximation for structured high-dimensional distributions via learning. *Advances in Neural Information Processing Systems*, 33:14699–14711, 2020.

Bhattacharyya, R. and Chakraborty, S. Property testing of joint distributions using conditional samples. *ACM Transactions on Computation Theory (TOCT)*, 10(4):1–20, 2018.

Bhattacharyya, R., Chakraborty, S., Pote, Y., Sarkar, U., and Sen, S. Testing self-reducible samplers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7952–7960, 2024.

Cai, W., Shi, T., Zhao, X., and Song, D. Are you getting what you pay for? auditing model substitution in llm apis. *arXiv preprint arXiv:2504.04715*, 2025.

Canonne, C. and Rubinfeld, R. Testing probability distributions underlying aggregated data. In *International Colloquium on Automata, Languages, and Programming*, pp. 283–295. Springer, 2014.

Canonne, C., Diakonikolas, I., Kane, D., and Stewart, A. Testing bayesian networks, 2020. URL https://arxiv.org/abs/1612.03156.

Canonne, C. L. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pp. 1–100, 2020.

Canonne, C. L. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

Canonne, C. L., Ron, D., and Servedio, R. A. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015.

Canonne, C. L., Chen, X., Kamath, G., Levi, A., and Waingarten, E. Random restrictions of high dimensional distributions and uniformity testing with subcube conditioning. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 321–336. SIAM, 2021.

Chakrabarty, D., Chen, X., Ristic, S., Seshadhri, C., and Waingarten, E. Monotonicity testing of high-dimensional distributions with subcube conditioning. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1019–1030, 2025.

Chakraborty, S., Fischer, E., Goldhirsh, Y., and Matsliah, A. On the power of conditional samples in distribution testing. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 561–580, 2013.

Chan, S.-O., Diakonikolas, I., Valiant, P., and Valiant, G. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1193–1203. SIAM, 2014.

Chen, L., Zaharia, M., and Zou, J. How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024.

Chu, X., Talluri, S., Lu, Q., and Iosup, A. An empirical characterization of outages and incidents in public services for large language models. In *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering, ICPE 2025*, pp. 69–80. ACM, 2025.

Clark, C. Provider variance: Introducing exacto. https://openrouter.ai/announcements/provider-variance-introducing-exacto, October 2025. OpenRouter Announcement.

Dutta, A., Krishnan, S., Kwatra, N., and Ramjee, R. Accuracy is not all you need. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Eyuboglu, S., Goel, K., Desai, A., Chen, L., Monfort, M., Ré, C., and Zou, J. Model changelists: Characterizing updates to ml models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2432–2453, 2024.

Fireworks AI. How fireworks evaluates quantization precisely and interpretably, 2024. URL https://fireworks.ai/blog/fireworks-quantization. Fireworks AI Blog.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Gao, I., Liang, P., and Guestrin, C. Model equality testing: Which model is this api serving? In *The Thirteenth International Conference on Learning Representations*, 2024.

Gibbs, A. L. and Su, F. E. On choosing and bounding probability metrics. *International statistical review*, 70 (3):419–435, 2002.

Giles, M. B. Multilevel monte carlo path simulation. *Operations research*, 56(3):607–617, 2008.

Giles, M. B. Multilevel monte carlo methods. *Acta numerica*, 24:259–328, 2015.

He, H. and Thinking Machines Lab. Defeating nondeterminism in LLM inference. *Thinking Machines Lab: Connectionism*, September 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Heinrich, S. Multilevel monte carlo methods. In *International Conference on Large-Scale Scientific Computing*, pp. 58–67. Springer, 2001.

Hooper, C., Kim, S., Mohammadzadeh, H., Mahoney, M. W., Shao, Y. S., Keutzer, K., and Gholami, A. Kvquant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

Li, Y., Wei, F., Zhang, C., and Zhang, H. EAGLE-2: faster inference of language models with dynamic draft trees. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pp. 7421–7432. Association for Computational Linguistics, 2024.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Liu, Z., Yuan, J., Jin, H., Zhong, S. H., Xu, Z., Braverman, V., Chen, B., and Hu, X. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Forty-first International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research, pp. 32332–32344. PMLR / OpenReview.net, 2024.

Meel, K. S., Kumar, G., and Pote, Y. Distance estimation for high-dimensional discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 955–963. PMLR, 2025.

Narayanan, S. On distribution testing in the conditional sampling model. *arXiv preprint arXiv:2007.09895*, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Valiant, G. and Valiant, P. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

Van Handel, R. Probability in high dimension. Technical report, Princeton University, 2014.

Zhou, Y., Lyu, K., Rawat, A. S., Menon, A. K., Rostamizadeh, A., Kumar, S., Kagy, J., and Agarwal, R. Distillspec: Improving speculative decoding via knowledge distillation. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

Zhu, X., Ye, Y., Qiu, T., Zhu, H., Tan, S., Mannan, A., Michala, J., Popa, R. A., and Neiswanger, W. Auditing black-box llm apis with a rank-based uniformity test. *arXiv preprint arXiv:2506.06975*, 2025.

A. Properties of TV Distance

Here we derive some equivalent formulations of the TV distance used in our estimation algorithms. We use the notation $(x)_+ := \max(x, 0)$.

Lemma 1 (Equivalent representations of total variation)**.** *Let $\pi, \mu \in \mathcal{P}(\Sigma^n)$, then $\mathrm{TV}(\pi, \mu)$ admits the following equivalent representations:*

$$\mathrm{TV}(\pi, \mu) = \mathbb{E}_{X\sim\pi}\left[\left(1 - \frac{\mu(X)}{\pi(X)}\right)_+\right], \tag{5}$$

$$\mathrm{TV}(\pi, \mu) = \mathbb{E}_{X\sim\frac{\pi+\mu}{2}}\left[\frac{|\pi(X) - \mu(X)|}{\pi(X) + \mu(X)}\right]. \tag{6}$$

Proof. By definition,

$$\mathrm{TV}(\pi, \mu) = \frac{1}{2}\sum_{x\in\Sigma^n} |\mu(x) - \pi(x)| = \sum_{x\in\Sigma^n} (\pi(x) - \mu(x))_+.$$

We first prove (5) by noting that

$$\mathbb{E}_{X\sim\pi}\left[\left(1 - \frac{\mu(X)}{\pi(X)}\right)_+\right] = \sum_{x\in\mathrm{supp}(\pi)} \pi(x)\left(1 - \frac{\mu(x)}{\pi(x)}\right)_+ = \sum_{x\in\Sigma^n} (\pi(x) - \mu(x))_+ = \mathrm{TV}(\pi, \mu).$$

A similar direct calculation yields (6):

$$\mathbb{E}_{X\sim\frac{\pi+\mu}{2}}\left[\frac{|\pi(X) - \mu(X)|}{\pi(X) + \mu(X)}\right] = \frac{1}{2}\sum_{x\in\Sigma^n} |\pi(x) - \mu(x)| = \mathrm{TV}(\pi, \mu).$$

□

B. TV Estimation via Prefix Logit Queries

In this section, we study Problem 1 under access to the prefix logit oracles $(\mathcal{O}^{\pi}_{\mathrm{logit}}, \mathcal{O}^{\mu}_{\mathrm{logit}})$. We first present our distance estimator and prove its query-complexity upper bound in Appendix B.1. The key observation is that exact prefix logit access simulates an exact log-probability oracle over full trajectories: for any $x \in \Sigma^n$,

$$\ell_\pi(x) := \log \pi(x_{1:n}) = \sum_{i=0}^{n-1} \log \pi(x_{i+1} \mid x_{1:i}), \tag{7}$$

and a similar expression holds for μ. Hence, density evaluations for π or μ cost n prefix logit oracle queries. We adapt the likelihood-ratio estimator of (Canonne & Rubinfeld, 2014), which is a bounded, unbiased estimator for $\mathrm{TV}(\pi, \mu)$ leveraging query access to $\frac{\mu}{\pi}$ via (5), achieving query complexity $O\left(\frac{n}{\varepsilon^2}\log\frac{1}{\delta}\right)$.

In Appendix B.2, we show that this reduction is tight in the worst case for any algorithm with prefix logit oracle access. In particular, our lower bound construction implies that, when $n \gg 1$, any algorithm solving Problem 1 with access to $(\mathcal{O}^{\pi}_{\mathrm{logit}}, \mathcal{O}^{\mu}_{\mathrm{logit}})$ requires $\Omega(\frac{n}{\varepsilon^2})$ queries.

B.1. Likelihood-ratio estimator

Our estimator in Algorithm 1 follows from the likelihood-ratio representation in equation (5): it is the empirical mean of a truncated likelihood-ratio functional evaluated on samples from one of the two distributions. To implement Algorithm 1, we need two primitives: (a) sampling a trajectory from one distribution, and (b) evaluating the probability of a given trajectory under either π or μ. Both primitives are available from accessing prefix logit oracles. Firstly, sampling can be performed sequentially by querying the prefix logit oracle, and drawing the next token appropriately. Moreover, given a trajectory $x_{1:n}$, its log-probability is computed via (7). Each primitive therefore requires n queries. Since the estimator averages bounded independent summands, standard concentration inequalities recalled in Theorem 4 give the desired high-probability guarantee.

Algorithm 1 LogitAccessTV($\mathcal{O}^{\pi}_{\text{logit}}, \mathcal{O}^{\mu}_{\text{logit}}, \varepsilon, \delta$)

$N \leftarrow \left\lceil \frac{\log \frac{2}{\delta}}{2\varepsilon^2} \right\rceil$
Generate $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \pi$ via $\mathcal{O}^{\pi}_{\text{logit}}$
for $i = 1, \ldots, N$ **do**
  $L_{\pi,i} \leftarrow \log \pi(X_i)$, $L_{\mu,i} \leftarrow \log \mu(X_i)$ {Evaluate using n prefix logit oracle queries via (7).}
  $R_i \leftarrow (1 - \exp(L_{\mu,i} - L_{\pi,i}))_+$
end for
$\hat{d} \leftarrow \frac{1}{N} \sum_{i=1}^{N} R_i$
return $\hat{d}$

Theorem 4. *Algorithm 1 solves Problem 1 with $O(\frac{n}{\varepsilon^2} \log \frac{1}{\delta})$ queries to $(\mathcal{O}^{\pi}_{\text{logit}}, \mathcal{O}^{\mu}_{\text{logit}})$.*

Proof. Letting $\log 0 := -\infty$, we obtain that $R_i = (1 - \frac{\mu(X_i)}{\pi(X_i)})_+$. By Lemma 1, we know that $\mathbb{E}[R_i] = \text{TV}(\pi, \mu)$. Moreover, since $\{R_i\}_{i \in [N]} \in [0, 1]$ are independent, Hoeffding's inequality gives

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} R_i - \text{TV}(\pi, \mu)\right| \geq \varepsilon\right] \leq 2\exp(-2N\varepsilon^2).$$

Taking N as stated gives the desired probability bound. Each sequence log-likelihood evaluation and sample generation uses n queries, which gives us $3nN = O(\frac{n}{\varepsilon^2} \log \frac{1}{\delta})$ query complexity bound. □

B.2. Lower bound

In this subsection, we take $\Sigma = \{0, 1\}$ to be a binary alphabet. We provide a hard instance showing that, for constant δ, Theorem 4 achieves the optimal query complexity up to a constant factor. Intuitively, this hard instance reduces to estimating a parameter $p \in (0, 1)$ up to error ε, in a way such that it takes $\approx n$ prefix logit queries to reveal one sample $\sim \text{Bern}(p)$.

Hard instance construction. Fix $r \in [1, n-2]$, let $t = n - r$, and set $L := |\Sigma^r| = 2^r$. We divide any $x \in \Sigma^n$ into two parts: $x = u \oplus z$, where $u \in \Sigma^r, z \in \Sigma^t$. The prefix u serves as a block index, while the suffix z specifies a leaf inside that block. In each block, we randomly choose whether $\mu_u, \pi_u \in \mathcal{P}(\Sigma^t)$ are identical or separated. If they are separated, they share the same $(t-1)$-bit prefix inside the block and differ only at the final bit, so the discrepancy takes t queries to reveal.

Formally, for $p \in [0, 1]$, define $(\mu, \pi) \sim \mathcal{D}_p$ as follows. For each $u \in \Sigma^r$, draw

$$X_u \overset{\text{i.i.d.}}{\sim} \text{Bern}(p), \qquad Z_u \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{0,1\}^{t-1}), \qquad B_u \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{0,1\}).$$

Let $a_u := Z_u \oplus B_u$, $b_u := Z_u \oplus (1 - B_u)$, be sibling leaves. Define distributions $\mu_u, \pi_u \in \mathcal{P}(\Sigma^t)$ as

$$(\mu_u, \pi_u) = \begin{cases} \left(\frac{1}{2}\delta_{a_u} + \frac{1}{2}\delta_{b_u}, \frac{1}{2}\delta_{a_u} + \frac{1}{2}\delta_{b_u}\right), & X_u = 0, \\ (\delta_{a_u}, \delta_{b_u}), & X_u = 1. \end{cases}$$

Finally, let $\mu(u \oplus z) := \frac{1}{L}\mu_u(z)$, $\pi(u \oplus z) := \frac{1}{L}\pi_u(z)$. A sample from μ is generated by drawing $u \sim \text{Unif}(\Sigma^r)$ and then drawing $z \sim \mu_u$; a sample from π is generated analogously using π_u.

This construction makes the TV distance easy to compute. Since different prefixes u define disjoint blocks, the global distance decomposes as the average of the suffix distances:

$$\text{TV}(\mu, \pi) = \frac{1}{L}\sum_{u \in \Sigma^r} \text{TV}(\mu_u, \pi_u) = \frac{1}{L}\sum_{u \in \Sigma^r} X_u. \tag{8}$$

The above display shows that, under $(\mu, \pi) \sim \mathcal{D}_p$, $\text{TV}(\mu, \pi)$ is exactly the empirical average of L i.i.d. Bernoulli variables. Hence, $\text{TV}(\mu, \pi)$ is close to p with high probability. Therefore, we can show that an accurate estimate of $\text{TV}(\mu, \pi)$ distinguishes the two hypotheses

$$H_0 : (\mu, \pi) \sim \mathcal{D}_{p_0}, \qquad H_1 : (\mu, \pi) \sim \mathcal{D}_{p_1},$$

whenever $p_1 - p_0$ is sufficiently large.

Lemma 2. *Let $0 \le p_0 < p_1 \le 1$ and set $\Delta := \frac{p_1 - p_0}{4}$. Suppose $\mathcal{A}$ estimates $\mathrm{TV}(\mu, \pi)$ within additive error Δ with probability at least $\frac{2}{3}$ for every fixed pair (μ, π). Then there is a tester distinguishing $\mathcal{D}_{p_0}$ from $\mathcal{D}_{p_1}$ with success probability at least $\frac{2}{3} - 2\exp(-2L\Delta^2)$.*

Proof. By (8) and Hoeffding's inequality, we know that for $(\mu, \pi) \sim \mathcal{D}_p$,

$$\mathbb{P}\left[|\mathrm{TV}(\mu, \pi) - p| > \Delta\right] \le 2\exp(-2L\Delta^2).$$

The tester runs $\mathcal{A}$ and outputs p_1 iff $\hat{d} \ge \frac{p_0 + p_1}{2}$. Conditioned on the success of the tester and $|\mathrm{TV}(\mu, \pi) - p| \le \Delta$, we have $|\hat{d} - p| \le 2\Delta = \frac{p_1 - p_0}{2}$. Therefore, if $p = p_0$, then $\widehat{d} \le \frac{p_0 + p_1}{2}$; if $p = p_1$, then $\widehat{d} \ge \frac{p_0 + p_1}{2}$. Hence the tester succeeds on these two events. Applying a union bound gives a success probability of at least $\frac{2}{3} - 2\exp(-2L\Delta^2)$. □

Lemma 2 shows it is enough to lower bound the oracle query complexity of estimating $\mathrm{TV}(\mu, \pi)$. Intuitively, we wish to show that many (u, X_u) must be revealed to perform this estimate.

Lemma 3 shows that with only B prefix queries, any adaptive algorithm can reveal at most $O(\frac{B}{t})$ such blocks u in expectation. We formalize this through the query-answer transcript. A possibly randomized adaptive algorithm chooses each query $W_i \in \Sigma^{<n}$ as a function of its private randomness and the previous query-answer pairs. We write

$$\mathsf{Tr} = \big((W_1, \mathcal{O}(W_1)), \ldots, (W_T, \mathcal{O}(W_T))\big)$$

for the full transcript, where $\mathcal{O}(W_i)$ is the oracle answer and $T \le B$ is the stopping time. For simplicity, we assume that $\mathcal{O}(W_i)$ reveals both $\mathcal{O}^{\pi}_{\mathrm{logit}}(W_i)$ and $\mathcal{O}^{\mu}_{\mathrm{logit}}(W_i)$, which is without loss of generality up to a constant factor in the oracle query complexity.

The crux of Lemma 3's argument is that X_u can affect the transcript only if the algorithm queries the hidden parent $u \oplus Z_u$. Hence, before learning anything about X_u, the algorithm must locate the hidden string $Z_u \in \{0,1\}^{t-1}$. We capture this notion formally by bounding the mutual information between the transcript and the initially-random set of strings $\mathcal{Z} = (Z_u)_{u \in \Sigma^r}$.

Lemma 3. *Consider any adaptive algorithm interacting with the oracle in Definition 1 on an instance (μ, π) drawn from $\mathcal{D}_p$. Suppose its worst-case query complexity is B. Let $\mathcal{G} \subseteq \Sigma^r$ be the set of $u \in \Sigma^r$ for which $W_j = u \oplus Z_u$ for some $j \le T$. There is a universal constant $C > 0$ such that*

$$\mathbb{E}[|\mathcal{G}|] \le C \cdot \frac{B}{t}.$$

Proof. Throughout, let q denote a random seed specifying internal randomness, which is drawn and fixed before interactions with $\mathcal{O}$. Thus, the algorithm is deterministic conditioned on q.

Let $\ell = t - 1 \ge \frac{t}{2}$ and let $\mathcal{Z} = (Z_u)_{u \in \Sigma^r}$. Inside any block $u \in \Sigma^r$, both μ_u and π_u are supported on the two sibling leaves $Z_u \oplus B_u$ and $Z_u \oplus (1 - B_u)$. Hence, before $u \oplus Z_u$ is queried, the oracle only reveals whether the queried prefix is consistent with Z_u and, if so, reveals the next bit of Z_u.[3] When the hidden $u \oplus Z_u$ is queried, the answer is $(\frac{1}{2}(e_0 + e_1), \frac{1}{2}(e_0 + e_1))$ if $X_u = 0$, and is $(\delta_{B_u}, \delta_{1-B_u})$ if $X_u = 1$. Therefore X_u can affect the oracle answer only at the query $u \oplus Z_u$.

Let Tr be the full transcript. We first upper bound the mutual information $I(\mathcal{Z}; \mathsf{Tr} \mid s)$. As shown above, each oracle answer lies in a constant-sized set so each query contributes at most $h_0 = O(1)$ bits of mutual information about $\mathcal{Z}$. Because the algorithm makes at most B queries, we have $I(\mathcal{Z}; \mathsf{Tr} \mid q) \le h_0 B$. Averaging over the internal randomness q, we obtain $I(\mathcal{Z}; \mathsf{Tr}) \le h_0 B$.

Conversely, if $u \in \mathcal{G}$, then the transcript contains a query of the form $u \oplus Z_u$, so it determines the full string Z_u. Thus, after observing the transcript, revealed blocks u have no remaining entropy in their hidden parents, while each unrevealed block u has at most ℓ bits of remaining entropy. Hence

$$H(\mathcal{Z} \mid \mathsf{Tr}) \le \ell\big(L - \mathbb{E}[|\mathcal{G}|]\big).$$

[3] If not, the model returns the distinguished symbol $(\perp, \perp)$, which is included in the transcript and carries only $O(1)$ bits.

Since the unconditional entropy $H(\mathcal{Z}) = L\ell$, we get

$$I(\mathcal{Z}; \mathsf{Tr}) = H(\mathcal{Z}) - H(\mathcal{Z} \mid \mathsf{Tr}) \geq \ell\, \mathbb{E}[|\mathcal{G}|].$$

Combining the upper and lower bounds gives

$$\mathbb{E}[|\mathcal{G}|] \leq \frac{h_0 B}{\ell} \leq 2h_0 \frac{B}{t},$$

where we used $\ell = t - 1 \geq \frac{t}{2}$. □

We are now ready to prove the lower bound. Lemma 2 shows that a TV estimator for our random instance can distinguish whether the underlying bias is p_0 or p_1. In the construction, this bias appears only through the variables $X_u \sim \mathrm{Bern}(p)$; conditioned on these variables, the remaining randomness in Z_u and B_u is independent of p.

Thus, given an unknown coin of bias $p \in \{p_0, p_1\}$, we can simulate the oracle for $(\mu, \pi) \sim \mathcal{D}_p$ by lazily drawing a coin sample only when the transcript reveals some X_u. By Lemma 3, a cost-B transcript reveals only $O(\frac{B}{t})$ such blocks in expectation. Hence a cost-B TV estimator would yield a coin distinguisher using only $O(\frac{B}{t})$ coin samples. Combining this result with classic sample complexity lower bounds for Bernoulli parameter estimation yields Theorem 5.

Theorem 5. *Fix any $\varepsilon \in (0, \frac{1}{8})$, and let $n = \Omega(\log \frac{1}{\varepsilon})$ for a sufficiently large constant. Any algorithm that solves Problem 1 with probability at least $\frac{2}{3}$ requires $\Omega(\frac{n}{\varepsilon^2})$ queries to $(\mathcal{O}^{\pi}_{\mathrm{logit}}, \mathcal{O}^{\mu}_{\mathrm{logit}})$.*

Proof. Set $p_0 = \frac{1}{2}, p_1 = \frac{1}{2} + 4\varepsilon$. By assumption we can set $r = \Omega(\log \frac{1}{\varepsilon})$ large enough so that $n \geq 2r$ and the error term in Lemma 2 satisfies $2\exp(-2L\varepsilon^2) \leq \frac{1}{100}$. Suppose that there is an estimator $\mathcal{A}$ with worst-case prefix-query cost B. By Lemma 2, $\mathcal{A}$ yields a distinguisher between

$$(\mu, \pi) \sim \mathcal{D}_{p_0} \qquad \text{and} \qquad (\mu, \pi) \sim \mathcal{D}_{p_1}$$

with success probability at least $\frac{2}{3} - \frac{1}{100} > \frac{3}{5}$.

We now simulate this distinguisher in an equivalent way, using an unknown coin of bias $p \in \{p_0, p_1\}$. The simulator samples all of the hidden Z_u and B_u using its own randomness, but does not sample X_u in advance. Whenever the distinguishing algorithm reveals an X_u, equivalently when some $W_i = u \oplus Z_u$ in the transcript, the simulator flips the unknown coin and stores it as X_u. By lazy sampling, the simulated transcript has exactly the same law as the true transcript under $\mathcal{D}_p$.

Let N be the number of coin samples used by the simulator. Since each coin sample corresponds to one revealed block, Lemma 3 implies $\mathbb{E}[N] \leq C\frac{B}{t}$. By Markov's inequality, we may truncate the simulation so that it uses at most $100C\frac{B}{t}$ coin samples while losing only $\frac{1}{100}$ of success probability. Thus, we obtain a tester distinguishing $\mathrm{Bern}(p_0)$ from $\mathrm{Bern}(p_1)$ with success probability $> \frac{3}{5}$ using $O(\frac{B}{t})$ coin samples. Classical results, e.g., (Canonne, 2022), imply that distinguishing $\mathrm{Bern}(p_0)$ from $\mathrm{Bern}(p_1)$ requires $\Omega\left(\frac{1}{\varepsilon^2}\right)$ samples, which implies $B = \Omega\left(\frac{t}{\varepsilon^2}\right) = \Omega\left(\frac{n}{\varepsilon^2}\right)$. □

Remark 1 (Extension to weaker prefix oracles)**.** *The same lower bound also applies to weaker access models, such as noisy prefix distribution or prefix sampler oracles. The only oracle-specific part of the proof is Lemma 3, where we show that a transcript of B prefix queries can reveal at most $O(B/t)$ hidden blocks in expectation. This step is information-theoretic: the bottleneck is locating the hidden parent $u \oplus Z_u$, before which the transcript contains no information about the block variable X_u. Since noisy distribution access and sampler access are no more informative than exact prefix-logit access, the same mutual-information bound on the revealed set $\mathcal{G}$ continues to hold. The remaining reduction from TV estimation to distinguishing $\mathrm{Bern}(p_0)$ from $\mathrm{Bern}(p_1)$ is unchanged, and therefore these weaker oracle models inherit the same $\Omega(n/\varepsilon^2)$ lower bound.*

C. TV Estimation via Noisy Prefix Distribution Queries

In this section, we study Problem 1 under access to the noisy prefix distribution oracles $(\mathcal{O}^{\pi}_{\mathrm{ndist}}, \mathcal{O}^{\mu}_{\mathrm{ndist}})$. We begin by motivating the noisy prefix distribution oracle definition in Definition 3 and justifying the claims from Section 2.2, particularly, that our results for Definition 3 with $\sigma^2 = k$ capture the prefix sampling oracle in Definition 2. We then present our basic estimator and its approximation guarantee in Appendix C.2, followed by an improved variance reduction argument based on multilevel Monte Carlo in Appendix C.3 that yields our final query complexity bound.

C.1. Motivation for the relative variance model

Here we give two basic justifications for the relative variance parameterization used in our noisy prefix distribution oracle (Definition 3). First, under this parameterization, Definition 3 subsumes the two oracle models introduced earlier: exact logit queries correspond to zero-noise sketches, while prefix-conditional samples correspond to random one-hot probability vectors with $\sigma^2 \le K$.

Lemma 4. *A prefix logit oracle (Definition 1) is a noisy prefix distribution oracle (Definition 3) with $\sigma^2 = 0$. Moreover, drawing $a \in \Sigma$ from a prefix sampler oracle (Definition 2) and outputting $e_a \in \mathcal{P}(\Sigma)$ is a noisy prefix distribution oracle with $\sigma^2 = K$.*

Proof. The first statement is trivial. For the second, consider any $\pi \in \mathcal{P}(\Sigma^n)$ and any prefix $s \in \Sigma^i$ with $i \in [n-1]$. Then letting $a \sim \mathcal{O}^{\pi}_{\mathrm{samp}}(s)$, we have $\mathbb{E}e_a = \pi_s$ by definition, and

$$\begin{aligned}\mathbb{E}_{a\sim\pi_s}\left[\chi^2\left(e_a\|\pi_s\right)\right] &= \mathbb{E}_{a\sim\pi_s}\left[\sum_{b\in\Sigma:[\pi_s]_b>0}\frac{(\mathbb{I}_{a=b}-\pi_s(b))^2}{\pi_s(b)}\right]\\ &= \mathbb{E}_{a\sim\pi_s}\left[\frac{(1-\pi_s(a))^2}{\pi_s(a)}+1-\pi_s(a)\right] = \mathbb{E}_{a\sim\pi_s}\left[\frac{1-\pi_s(a)}{\pi_s(a)}\right] \le K.\end{aligned}$$

□

Second, Definition 3 also captures a simple form of noisy logit access with linear noise. Since language models produce logits before applying the softmax, a natural abstraction of inference-time instability is as an additive perturbation to the logits. Lemma 5 shows that small (additive) Gaussian logit perturbations induce bounded relative variance at the distributional level.

Lemma 5. *Fix $\ell \in \mathbb{R}^m$, let $s(\xi) := \mathrm{softmax}(\ell+\xi) \in \mathcal{P}(\Sigma)$, and let $s_a(\xi) = \mathrm{softmax}(\ell+\xi)_a$, for all $a \in \Sigma$. Let $p := \mathbb{E}_\xi[s(\xi)]$, where $\xi \sim N(0,\gamma^2 I_m)$ is an isotropic Gaussian. If $\gamma^2 < \frac{1}{2}$, then*

$$V := \mathbb{E}_{\xi\sim N(0,\gamma^2 I_m)}\left[\sum_{a\in\Sigma}\frac{(s_a(\xi)-p_a)^2}{p_a}\right] \le \frac{2\gamma^2}{1-2\gamma^2}.$$

Proof. By the Gaussian Poincaré inequality (Corollary 2.17, (Van Handel, 2014)), for any smooth (infinitely-differentiable) function f on $\mathbb{R}^m$, if $\xi \sim N(0,\gamma^2 I_m)$,

$$\mathrm{Var}[f(\xi)] \le \gamma^2\mathbb{E}\|\nabla f(\xi)\|_2^2.$$

Applying this fact to the scalar function $f(\xi) = s_a(\xi)$ for each $a \in \Sigma$, and using the fact that $\nabla_\xi s_a(\xi) = s_a(e_a - s)$ with $\|e_a - s\|_2^2 \le 2$, gives $\mathrm{Var}[s_a(\xi)] \le 2\gamma^2\mathbb{E}\left[s_a^2(\xi)\right]$. Thus,

$$V = \sum_{a\in\Sigma}\frac{\mathrm{Var}[s_a(\xi)]}{p(a)} \le 2\gamma^2\sum_{a\in\Sigma}\frac{\mathbb{E}\left[s_a^2(\xi)\right]}{p(a)} = 2\gamma^2(V+1).$$

We conclude the proof by rearranging the above inequality. □

C.2. Estimator construction

In this section, we develop a basic estimator for solving Problem 1 under the noisy access afforded by Definition 3. A naïve application of this estimator yields a suboptimal $\approx \frac{n^2\sigma^2}{\varepsilon^4}$ query complexity, which is improved by a variance reduction scheme in Section C.3.

To begin, we recall the mixture representation of the TV distance (6) from Lemma 1:

$$\begin{aligned}&\mathrm{TV}(\pi,\mu) = \mathbb{E}_{X\sim M}[Z(X)], \qquad M = \frac{\pi+\mu}{2},\\ &Z_{\pi,\mu}(x) := \begin{cases}\dfrac{|\pi(x)-\mu(x)|}{\pi(x)+\mu(x)}, & \pi(x)+\mu(x) > 0,\\ 0, & \pi(x)=\mu(x)=0.\end{cases}\end{aligned} \tag{9}$$

When (π, μ) is clear from context, we write $Z(x)$ instead of $Z_{\pi,\mu}(x)$. Our strategy is to construct an empirical approximation $\widehat{Z}$ to Z and then bound the expected squared error $\mathbb{E}_{X\sim M}[(\widehat{Z}(X) - Z(X))^2]$.

As a first step, we explain how to construct an approximation to each of the quantities $\pi(x)$ and $\mu(x)$ required by Z. Fix some $r \in \mathbb{N}$ and prefix $s \in \Sigma^{<n}$. We define a random estimator

$$\widehat{p}_{s,r} := \frac{1}{r}\sum_{j=1}^{r} Y^{\pi}_{s,j} \in \mathcal{P}(\Sigma), \qquad Y^{\pi}_{s,j} \overset{\text{i.i.d.}}{\sim} \mathcal{O}^{\pi}_{\text{ndist}}(s). \tag{10}$$

In other words, $\widehat{p}_{s,r}$ simply queries the noisy prefix distribution oracle $\mathcal{O}^{\pi}_{\text{ndist}}(s)$ for r times in total, and averages the outputs. For any distribution $\pi \in \mathcal{P}(\Sigma^n)$ and any $r \in \mathbb{N}$, we can therefore define a corresponding empirical distribution $\widehat{\pi}_r$ as follows: for a string $x \in \Sigma^n$, we let

$$\widehat{\pi}_r(x) := \prod_{i=0}^{n-1} \widehat{p}_{x_{1:i},r}(x_{i+1}) \tag{11}$$

be a random estimator for $\pi(x)$, where each of the $\widehat{p}_{x_{1:i},r}(x_{i+1})$ is estimated via (10) using r independent queries to $\mathcal{O}_{\text{ndist}}(x_{1:i})$. Thus, one evaluation of $\widehat{\pi}_r$ requires nr queries to $\mathcal{O}_{\text{ndist}}$.

We first bound the expected χ^2 divergence between $\pi \in \mathcal{P}(\Sigma^n)$ and its empirical counterpart $\widehat{\pi}_r$, where expectations are over the randomness defining $\widehat{\pi}_r$ (e.g., in the oracle queries).

Lemma 6. *For any $\pi \in \mathcal{P}(\Sigma^n)$, let $\widehat{\pi}_r$ be defined in (11), where for each prefix $s \in \Sigma^{<n}$, we generate an estimate $\widehat{p}_{s,r}$ as in (10) independently. Then if $\mathcal{O}_{\text{ndist}}$ is a noisy prefix distribution oracle (Definition 3) with relative variance σ^2, there is a universal constant C such that for any $r \geq Cn\sigma^2$,*

$$\mathbb{E}\left[\chi^2(\widehat{\pi}_r \| \pi)\right] \leq \frac{Cn\sigma^2}{r}.$$

Proof. We observe that almost surely $\widehat{\pi}_r \ll \pi$, since any token with zero conditional probability under any $\pi(\cdot \mid s)$ receives zero mass under $\widehat{p}_{s,r}$. Hence

$$1 + \chi^2(\widehat{\pi}_r \| \pi) = \mathbb{E}_{X\sim\pi}\left[\left(\frac{\widehat{\pi}_r(X)}{\pi(X)}\right)^2\right],$$

where the expectation also includes the randomness of querying $\mathcal{O}_{\text{ndist}}$. For any fixed prefix $s \in \Sigma^{<n}$,

$$\mathbb{E}_{a\sim\pi_s}\left[\left(\frac{\widehat{p}_{s,r}(a)}{\pi_s(a)}\right)^2\right] = \sum_{a:\pi_s(a)>0} \frac{\mathbb{E}[\widehat{p}_{s,r}(a)^2]}{\pi_s(a)} = \sum_{a:\pi_s(a)>0}\left(\pi_s(a) + \frac{\text{Var}[\widehat{p}_{s,r}(a)]}{\pi_s(a)}\right) \leq 1 + \frac{\sigma^2}{r}.$$

Applying this one-step estimate sequentially along the path $X_1, \ldots, X_n$, and using independence of the estimators $\widehat{p}_{s,r}$ across prefixes s, gives the claim:

$$\mathbb{E}_{X\sim\pi}\left[\left(\frac{\widehat{\pi}_r(X)}{\pi(X)}\right)^2\right] \leq \left(1 + \frac{\sigma^2}{r}\right)^n \leq 1 + \frac{Cn\sigma^2}{r}.$$

□

In implementing the estimator $\widehat{\pi}_r$ in (11), it is helpful to assume that in the background, a data structure has queried $\mathcal{O}_{\text{ndist}}(s)$ for r times for *every prefix* $s \in \Sigma^{<n}$. When our algorithm requires an estimate $\widehat{\pi}_r(x)$, the data structure lazily provides the estimates $\widehat{p}_{s,r}$ for each $s = x_{1:i}$, and we use these estimates to compute $\widehat{\pi}_r(x)$. We similarly define an estimator $\widehat{\mu}_r$ for each $r \in \mathbb{N}$.

We are now ready to define our empirical approximation $\widehat{Z}$ to Z in (9). For fixed $r \in \mathbb{N}$, define $\widehat{\pi}_r$ and $\widehat{\mu}_r$ as in (11), which we have query access to. We then let

$$\widehat{Z}(x) := \frac{|\widehat{\pi}_r(x) - \widehat{\mu}_r(x)|}{\widehat{\pi}_r(x) + \widehat{\mu}_r(x)} \tag{12}$$

for any x where $\pi(x) + \mu(x) > 0$. If we observe $\widehat{\pi}_r(x) + \widehat{\mu}_r(x) = 0$, we set the corresponding $\widehat{Z}(x) = 0$.

In Lemma 7 and 8, we reduce bounding the squared error between $\widehat{Z}$ and Z to understanding the χ^2 divergences between $(\widehat{\pi}_r, \pi)$ and $(\widehat{\mu}_r, \mu)$.

Lemma 7. *For $u, v \geq 0$, let $\phi(u, v) := \frac{|u-v|}{u+v}, \forall u, v \geq 0, u + v > 0$, and $\phi(0, 0) = 0$. Then for all $a, b, c, d \geq 0$ with $a + b > 0$,*

$$|\phi(c, d) - \phi(a, b)| \leq 4 \left(\frac{|c - a| + |d - b|}{a + b} \right).$$

Proof. Let $s := a + b$, $t := c + d$, and $\Delta := |c - a| + |d - b|$. If $t \geq \frac{s}{2}$, then

$$\begin{aligned} |\phi(c, d) - \phi(a, b)| &\leq \left| \frac{c - d}{t} - \frac{a - b}{s} \right| \\ &\leq \frac{|(c - d) - (a - b)|}{t} + |a - b| \frac{|t - s|}{st} \\ &\leq \frac{\Delta}{t} + \frac{\Delta}{t} \leq \frac{4\Delta}{s}. \end{aligned}$$

We conclude the proof by noting that when $t < \frac{s}{2}$, $\Delta > \frac{s}{2}$, while $|\phi(c, d) - \phi(a, b)| \leq 1$. □

Lemma 8. *Let $\mathcal{O}^{\pi}_{\text{ndist}}$, $\mathcal{O}^{\mu}_{\text{ndist}}$ respectively be noisy prefix distribution oracles with relative variance σ^2 for $\pi, \mu \in \mathcal{P}(\Sigma^n)$. There is a universal constant $C > 0$ such that, for every $r \geq Cn\sigma^2$, defining M, Z as in (9) and $\widehat{Z}$ as in (12),*

$$\mathbb{E}_{X \sim M} \left[\left(\widehat{Z}(X) - Z(X) \right)^2 \right] \leq \frac{Cn\sigma^2}{r}.$$

Proof. Observe that for $X \sim M$, we must have $\pi(X) + \mu(X) > 0$ by definition. Applying Lemma 7 with $(a, b, c, d) \leftarrow (\pi(X), \mu(X), \widehat{\pi}_r(X), \widehat{\mu}_r(X))$ shows that

$$|\widehat{Z}(X) - Z(X)| \leq 4 \left(\frac{|\widehat{\pi}_r(X) - \pi(X)| + |\widehat{\mu}_r(X) - \mu(X)|}{\pi(X) + \mu(X)} \right).$$

Therefore, using $(a + b)^2 \leq 2a^2 + 2b^2$ and $M = \frac{\pi+\mu}{2}$,

$$\begin{aligned} \mathbb{E}_{X \sim M} \left[(\widehat{Z}(X) - Z(X))^2 \right] &\leq 16 \sum_{x \in \Sigma^n} \frac{(\widehat{\pi}_r(x) - \pi(x))^2 + (\widehat{\mu}_r(x) - \mu(x))^2}{\pi(x) + \mu(x)} \\ &\leq 16\chi^2(\widehat{\pi}_r \| \pi) + 16\chi^2(\widehat{\mu}_r \| \mu). \end{aligned}$$

Taking expectations over the randomness of $\widehat{\pi}_r$, $\widehat{\mu}_r$ and applying Lemma 6 proves the claim. □

Lemma 8 already implies a nontrivial algorithm for solving Problem 1. In particular, taking $r = O(\frac{n\sigma^2}{\varepsilon^2})$ in Lemma 8 implies that $\widehat{Z}$ has bias ε with respect to Z, and each evaluation of $\widehat{Z}$ requires $O(nr)$ queries. Combining this calculation with the $O(\frac{1}{\varepsilon^2})$ evaluations of $\widehat{Z}$ required to estimate its expectation to ε additive error gives a total query complexity of

$$O\left(\frac{n^2\sigma^2}{\varepsilon^2} \right) \cdot O\left(\frac{1}{\varepsilon^2} \right) = O\left(\frac{n^2\sigma^2}{\varepsilon^4} \right).$$

This bound already improves upon the $O(\frac{n^3}{\varepsilon^5})$ queries required by the prior work of (Meel et al., 2025) in the regime $m = |\Sigma| = O(1)$. In Section C.3, we show how to use the squared error bound in Lemma 8 in a black-box way within a variance reduction scheme to improve this complexity bound.

Remark 2 (Asymmetric noise)**.** *Lemma 8 generalizes straightforwardly to settings where there is asymmetry in the relative variance bounds for $\mathcal{O}^{\pi}_{\text{ndist}}$, $\mathcal{O}^{\mu}_{\text{ndist}}$. In particular, if these bounds are respectively σ_π^2, σ_μ^2, and we instead use $r_\pi \in \mathbb{N}$ oracle queries at each prefix to define $\widehat{\pi}_{r_\pi}$ (similarly, $r_\mu \in \mathbb{N}$ queries to define $\widehat{\mu}_{r_\mu}$), it is simple to extend Lemma 8 to yield an expected squared error of*

$$\frac{Cn\sigma_\mu^2}{r_\mu} + \frac{Cn\sigma_\pi^2}{r_\pi}.$$

This asymmetry could be useful in saving samples when, e.g., we have exact query access to one of the autoregressive models (π, μ), but we have noisy or sample access to the other. To ease notational burden, we assume $\sigma_\mu = \sigma_\pi$ for simplicity henceforth.

C.3. Variance reduction through multilevel Monte Carlo

In this subsection, we improve upon a naïve application of Lemma 8 by estimating a telescoping sequence of increasingly-accurate approximations in a *multilevel Monte Carlo* (MLMC) variance reduction scheme. The analysis of our variance-reduced scheme uses Lemma 8 in a black-box way.

Fix an integer $L = O(\log \frac{1}{\varepsilon})$ in the following discussion, and for all levels $\ell = 0, 1, \ldots, L$, define

$$r_\ell := 1 + \lceil cn\sigma^2\, 2^\ell \rceil\,, \tag{13}$$

where c is a large constant and σ parameterizes the relative variance of $\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}}$. The key to our variance reduction is to couple the external draw $X \sim M$ in adjacent levels, so that the correction from level $\ell - 1$ to level ℓ has variance proportional to its approximation error.

More concretely, for all $\ell = 0, 1, \ldots, L$, let Z_ℓ denote a random independent copy of the estimator $\widehat{Z}$ in (12), with $r \leftarrow r_\ell$ as set in (13). Also, define

$$\theta_\ell := \mathbb{E}[Z_\ell(X)], \qquad \theta := \mathrm{TV}\,(\pi, \mu) = \mathbb{E}_{X\sim M}\left[Z(X)\right],$$

where whenever we take expectations over Z_ℓ, we also include the randomness from the noisy queries to $\mathcal{O}_{\text{ndist}}$. By Lemma 8, choosing a large enough c yields

$$\mathbb{E}\left[(Z_\ell(X) - Z(X))^2\right] \le C2^{-\ell}.$$

Therefore, by the Cauchy-Schwarz inequality,

$$|\theta_\ell - \theta| = |\mathbb{E}[Z_\ell(X) - Z(X)]| \le \sqrt{\mathbb{E}[(Z_\ell(X) - Z(X))^2]} \le C2^{-\frac{\ell}{2}}, \tag{14}$$

so that if $L = O(\log \frac{1}{\varepsilon})$ is large enough, $|\theta_L - \theta| \le \frac{\varepsilon}{2}$. Our goal is thus to estimate θ_L to $\frac{\varepsilon}{2}$ additive error. To do so, we write it as a telescoping sum:

$$\theta_L = \theta_0 + \sum_{\ell=1}^{L} (\theta_\ell - \theta_{\ell-1}). \tag{15}$$

Our goal is now to design estimators for each element $\theta_\ell - \theta_{\ell-1}$ within the sum (15). To do so, we couple the outer draw $X \sim M$ used to define these estimators. In particular, define $Z_{-1}(\cdot) \equiv 0$ for simplicity, and for each $\ell = 0, 1, \ldots, L$, define an estimator

$$Y_\ell := Z_\ell(X) - Z_{\ell-1}(X), \text{ where } X \sim M. \tag{16}$$

It is clear that this construction yields

$$\mathbb{E}[Y_0] = \theta_0, \qquad \mathbb{E}[Y_\ell] = \theta_\ell - \theta_{\ell-1} \text{ for } \ell = 0, 1, \ldots, L.$$

We now bound the variance of each estimator Y_ℓ.

Lemma 9. *For every $\ell = 0, \ldots, L$, we have $\mathrm{Var}[Y_\ell] \le \min(4, C2^{-\ell})$ for a universal constant $C > 0$. Moreover, we can obtain an i.i.d. copy of Y_ℓ using $O(n + n^2\sigma^2 2^\ell)$ queries to $\mathcal{O}^\pi_{\text{ndist}}$ and $\mathcal{O}^\mu_{\text{ndist}}$.*

Proof. The bound $\mathrm{Var}[Y_\ell] \le 4$ is immediate because each $Z_\ell(X)$ lies in $[0, 1]$ by definition, so each Y_ℓ lies in $[-1, 1]$. For the remaining bound, letting $\ell \ge 1$,

$$Y_\ell = Z_\ell(X) - Z_{\ell-1}(X) = (Z_\ell(X) - Z(X)) + (Z(X) - Z_{\ell-1}(X)), \tag{17}$$

where $X \sim M$. Thus, applying $(a + b)^2 \le 2a^2 + 2b^2$,

$$\mathrm{Var}[Y_\ell] \le \mathbb{E}[Y_\ell^2] \le 2\mathbb{E}[(Z_\ell(X) - Z(X))^2] + 2\mathbb{E}[(Z_{\ell-1}(X) - Z(X))^2] \le C2^{-\ell},$$

where the last step follows from adjusting the constant C in Lemma 8. The query complexity is immediate from the definition of $\widehat{Z}$ and our sample bound r_ℓ in (13). □

The proof of Lemma 9 reveals why it is important that the same $X \sim M$ is used to define Z_ℓ and $Z_{\ell-1}$ in the coupled increment Y_ℓ. Indeed, if different samples were used, then the identity (17) would not hold and we would not inherit the separate squared error bounds from Lemma 8.

Lemma 9 shows that the coupled increments Y_ℓ have geometrically-decaying variance across levels. This suggests a non-uniform sampling strategy, using more i.i.d. draws of Y_ℓ on coarse levels (small ℓ), where the variance is larger, and fewer samples on fine levels (large ℓ), where each sample is more expensive. By balancing the per-level variance contribution against the per-level sampling cost, we obtain an MLMC estimator that reduces the total query complexity.

We now explain how to allocate samples across levels. Let

$$N_\ell := \left\lceil A \cdot \frac{L+1}{\varepsilon^2} \cdot 2^{-\ell} \right\rceil, \qquad \ell = 0, 1, \ldots, L, \tag{18}$$

where $A > 0$ is a universal constant. Let $Y_{\ell,1}, \ldots, Y_{\ell,N_\ell}$ be independent copies of Y_ℓ, and define

$$T := \frac{1}{N_0} \sum_{j=1}^{N_0} Y_{0,j} + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{j=1}^{N_\ell} Y_{\ell,j}. \tag{19}$$

Our final estimator simply takes the median of $\log(\frac{1}{\delta})$ copies of T defined in (19). We provide pseudocode for the overall method in Algorithms 2 and 3, and analyze its correctness in Theorem 6.

Algorithm 2 $\mathsf{CoupledIncrement}(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}}, \ell, \sigma, n)$

$X \sim M = \frac{1}{2}(\pi + \mu)$ {Uses n queries to $\mathcal{O}^\nu_{\text{ndist}}$ for $\nu \sim_{\text{unif.}} \{\mu, \pi\}$, as $\mathcal{O}^\nu_{\text{ndist}}$ simulates $\mathcal{O}^\nu_{\text{samp}}$}
$Z_\ell(X) \leftarrow$ copy of $\widehat{Z}$ in (12) with $r \leftarrow r_\ell$ in (13) {Uses nr_ℓ queries to $(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}})$}
if $\ell = 0$ **then**
  $Z_{\ell-1}(X) \leftarrow 0$
else
  $Z_{\ell-1}(X) \leftarrow$ copy of $\widehat{Z}$ in (12) with $r \leftarrow r_{\ell-1}$ in (13) {Uses $nr_{\ell-1}$ queries to $(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}})$}
end if
return $Z_\ell(X) - Z_{\ell-1}(X)$

Algorithm 3 $\mathsf{NoisyDistAccessTV}(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}}, \varepsilon, \delta, \sigma, n)$

$R \leftarrow \lceil 10 \log \frac{1}{\delta} \rceil$
$L \leftarrow \lceil C + 2 \log_2 \frac{1}{\varepsilon} \rceil$
for $s = 1, \ldots, R$ **do**
  $T_s \leftarrow 0$ {Running sum for s^{th} evaluation of T in (19)}
  for $\ell = 0, \ldots, L$ **do**
    $S_\ell \leftarrow 0$
    for $j = 1, \ldots, N_\ell$ **do**
      $Y_{\ell,j} \leftarrow \mathsf{CoupledIncrement}(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}}, \ell, \sigma, n)$
      $S_\ell \leftarrow S_\ell + Y_{\ell,j}$
    end for
    $T_s \leftarrow T_s + \frac{S_\ell}{N_\ell}$ where N_ℓ is as in (18)
  end for
end for
return $\text{median}\{T_1, \ldots, T_R\}$

We are now ready to prove Theorem 3, stated slightly more formally here:

Theorem 6. *If $(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}})$ are respectively noisy prefix distribution oracles with relative variance σ^2 for (π, μ), Algorithm 3 solves Problem 1 with $O(\frac{n+n^2\sigma^2}{\varepsilon^2} \log^2(\frac{1}{\varepsilon}) \log(\frac{1}{\delta}))$ queries to $(\mathcal{O}^\pi_{\text{ndist}}, \mathcal{O}^\mu_{\text{ndist}})$.*

Proof. We follow the notation in Algorithm 3. Observe that for any $s = 1, \ldots, R$, we have $\mathbb{E}[T_s] = \theta_L$ by using the calculation (15). Thus, for a large enough constant in the choice of L, (14) implies $|\mathbb{E}[T_s] - \theta| \leq \frac{\varepsilon}{2}$. Moreover, by independence across calls to CoupledIncrement, we have by applying the variance bound in Lemma 9 that

$$\mathrm{Var}[T_s] = \sum_{\ell=0}^{L} \frac{\mathrm{Var}[Y_\ell]}{N_\ell} \leq \frac{4}{N_0} + \sum_{\ell=1}^{L} \frac{C2^{-\ell}}{N_\ell}.$$

Using N_ℓ as defined in (18), each summand is at most $O(\frac{\varepsilon^2}{A(L+1)})$. Hence, by choosing A sufficiently large, we obtain that $\mathrm{Var}[T_s] \leq \frac{\varepsilon^2}{12}$. Chebyshev's inequality then gives the single-estimate bound

$$\mathbb{P}\left[|T_s - \mathrm{TV}(\pi, \mu)| > \varepsilon\right] \leq \mathbb{P}\left[|T_s - \theta_L| > \frac{\varepsilon}{2}\right] \leq \frac{4\mathrm{Var}[T_s]}{\varepsilon^2} \leq \frac{1}{3}.$$

Finally, a standard application of Hoeffding's inequality shows that the median of R independent copies of an estimator T obeying the above bound lies within ε of $\theta = \mathrm{TV}(\pi, \mu)$ with probability at least $1 - \delta$. The query complexity follows from Lemma 9 with our definition (18):

$$\sum_{\ell=0}^{L} RN_\ell \cdot O(n + n^2\sigma^2 2^\ell) = O\left(\frac{(n + n^2\sigma^2)(L+1)^2 \log \frac{1}{\delta}}{\varepsilon^2}\right) + O\left((n + n^2\sigma^2)2^L \log \frac{1}{\delta}\right),$$

where we conclude by using that $2^L = O(\frac{1}{\varepsilon^2})$ by our choice of L. □

As a corollary of Theorem 6 and the oracle reduction in Lemma 4, we obtain a query complexity bound for Problem 1 under prefix sampling oracle access.

Corollary 1. *Algorithm 3 solves Problem 1 with* $O(\frac{n^2 K}{\varepsilon^2} \log^2(\frac{1}{\varepsilon}) \log \frac{1}{\delta})$ *queries to* $(\mathcal{O}^{\pi}_{\mathrm{samp}}, \mathcal{O}^{\mu}_{\mathrm{samp}})$.

D. Experimental details

This appendix gives the reproducibility setup, an exact-logit sanity check on the synthetic instance, and the methodology behind the cross-engine measurement of Section 4.2: the replay scoring oracle and the support-mismatch decomposition of Figure 2(a).

D.1. Setup and instance parameters

Synthetic hard instance. We use the binary autoregressive lower-bound construction of Appendix B.2: each sequence factors as $x = u \oplus z$ with $u \in \Sigma^r$ a block index and $z \in \Sigma^{n-r}$ a within-block leaf, and within each active block the discrepancy between π and μ is concentrated at the final bit. Throughout the synthetic experiments we fix $n = 128$, $r = 12$, $p_{\mathrm{active}} = 0.4$, $\alpha = 0.49$, giving $\mathrm{TV}(\pi, \mu) = 0.388$ by direct enumeration. The noisy oracle of Definition 3 is realized by adding zero-mean Gaussian noise of standard deviation $\sigma\sqrt{\hat{p}(1-\hat{p})}$ to each realized next-token probability $\hat{p}$ before taking logarithms (so deterministic transitions stay deterministic and the K-average is unbiased for the true probability); at $\sigma = 0$ this reduces to the exact prefix logit oracle of Definition 1.

Cross-engine sampling protocol. We use `vllm 0.19.1` with `enforce_eager=True` and prefix caching disabled, and `sglang 0.5.10.post1` with CUDA graph capture disabled, both running `Qwen3-0.6B` at bf16, top-$k = 20$, temperature 1, with EOS suppression for fixed-length sampling. The shared input is the system prompt `"You are a helpful assistant."` followed by a user prompt; the reference prompt is `"Tell me a story about a robot learning to paint."`, which the Qwen3 chat template tokenizes to a 30-token prefix. For each side we sample $N = 512$ length-$n = 500$ continuations and score each prefix $r = 8$ times per engine with the replay oracle below; the π-side uses base seed 0 and the μ-side base seed 1, with each repeat using an independent per-call seed offset. Sampling and scoring of one engine occur inside a single engine session to avoid kernel-reload byte-identical artifacts on the r-axis. The cross-prompt experiment of Figure 2(b) repeats the identical protocol for six prompts: the reference (story) plus `"Explain how a refrigerator works to a curious ten-year-old."` (explanation), `"What were the main causes of the French Revolution?"` (history), `"Summarize the plot of Romeo and Juliet in three sentences."` (literature), 用中文解释什么是梯度下降。 (Chinese), and `"Write a Python function that returns the n-th Fibonacci number."` (code).

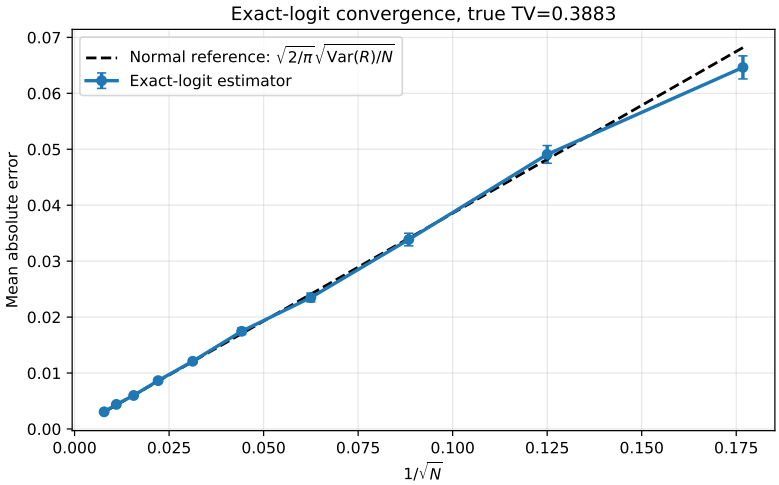

Figure 5. Mean absolute error of Algorithm 1 on the synthetic hard instance at $\sigma = 0$ (exact logits), against $1/\sqrt{N}$, with $\mathrm{TV}(\pi, \mu) = 0.388$ known by enumeration. Dashed line is the Gaussian-asymptotic reference $\sqrt{2/\pi}\,\sqrt{\mathrm{Var}(R)/N}$. The fitted slope is -0.493 vs the theoretical -0.500.

Replay scoring oracle. A faithful black-box oracle must score the engine's own decode path. Given a sampled trajectory X, we register a per-request logit processor that forces each step to the corresponding token of X (teacher-forcing), so the engine re-decodes along X under the sampling-time batch shape; the processor captures that step's top-k log-softmax *before* the forcing is applied — the engine's returned log-probabilities are computed after forcing and are contaminated, so the raw values must be captured inside the processor — and renormalizes over the top-k set to give one unbiased draw of the next-token distribution, with r replays giving r i.i.d. draws. We validate the oracle per engine: forcing is exact (zero own-zero, i.e. no trajectory on which an engine assigns zero r-averaged mass to its own sampled token; under the alternative whole-sequence `sglang` log-probability API this rate is 37%, which biases the estimate), and repeats are bit-identical for `sglang` and within $\approx 4 \times 10^{-5}$ of the sampling-time read-out for `vllm` — giving the per-engine self-noise ($\sigma_{\text{sglang}} = 0$, $\sigma_{\text{vllm}} \approx 8 \times 10^{-4}$) reported in Section 4.2.

Hardware and software. All experiments were run on a single NVIDIA GH200 Grace Hopper node with 96 GB HBM3e memory on a shared SLURM-managed cluster. The toolchain is `gcc 14.2.0` / `cuda 12.8` / `python 3.11.8`. The two engines use isolated virtual environments: vllm pairs with `torch 2.10.0+cu128` and `transformers 5.7.0`; sglang pairs with `torch 2.9.1+cu128` and `transformers 5.3.0`. Synthetic experiments run on a single CPU using NumPy with master seed 10.

D.2. Exact-logit sanity check

At $\sigma = 0$, Algorithm 1 reduces to a standard outer-trajectory Monte Carlo on the bounded estimator $R = (1 - \exp(L_\mu - L_\pi))_+$, with $\mathbb{E}[R] = \mathrm{TV}(\pi, \mu)$ by Lemma 1. Figure 5 plots the empirical mean absolute error against $1/\sqrt{N}$ and overlays the Gaussian-asymptotic reference $\sqrt{2/\pi}\,\sqrt{\mathrm{Var}(R)/N}$. The fitted slope is -0.493 against the theoretical -0.500, and the empirical curve sits slightly below the reference because $R \in [0, 1]$ has shorter-than-Gaussian tails. This confirms that the synthetic instance, the true-TV computation, and the estimator are mutually consistent before any noise is added.

D.3. Decomposition of the cross-engine estimator

The decomposition in Figure 2(a) splits the estimator output at each truncated length n' into two parts. The *support-mismatch contribution* at n' is the fraction of trajectories that contain at least one such cell within their first n' positions, where a *support-mismatch cell* is one in which, after r-averaging, one engine's empirical top-k assigns positive mass to the sampled token while the other does not. By construction, each such trajectory contributes the bounded value 1 from (12): the formula $|\widehat{\pi}_r(x) - \widehat{\mu}_r(x)|/(\widehat{\pi}_r(x) + \widehat{\mu}_r(x))$ evaluates to 1 whenever exactly one of $\widehat{\pi}_r, \widehat{\mu}_r$ is zero on the trajectory. The *shared-support log-probability-difference contribution* is the residual after subtracting the support-mismatch term, computed directly on

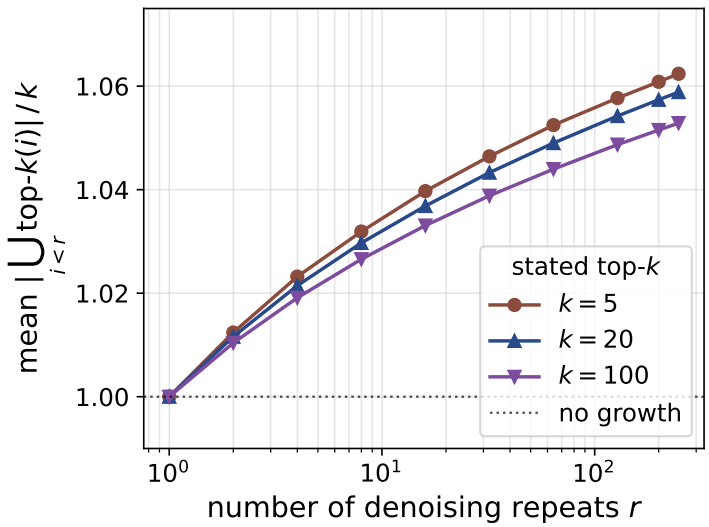

Figure 6. Mean empirical top-k support union size $|\bigcup_{i<r} \text{top-}k(i)|$ relative to the stated k, as a function of the number of repeated queries r at a fixed prefix on `Qwen3-0.6B`. Even at $r = 248$ the union exceeds the stated k by at most $\approx 7\%$, well below $2k$.

cells where both engines emit positive r-averaged probability and accumulating $\log \widehat{\pi}_r(x_t \mid x_{<t}) - \log \widehat{\mu}_r(x_t \mid x_{<t})$ across t. The support-mismatch term grows from 0% at $n' = 10$ to 29% at $n' = 500$ (and continues to $\approx 48\%$ by $n' = 2000$); the shared-support term saturates near $n' \approx 200$. The two cross over at $n' \approx 300$, and at $n' = 500$ the support-mismatch term contributes 49% of the total.

D.4. Why the multilevel scheme is not needed on this instance

The replay oracle is near-deterministic: its single-repetition self-noise floor is 0 for `sglang` and 0.007 for `vllm` at $n = 500$ (Section 4.2), orders of magnitude below the cross-engine $\mathrm{TV}(\pi, \mu) = 0.586$. The instance therefore sits in the logit-access regime of Theorem 2: the variance ratio $\rho_1 \approx 0$, a single level already attains the matched-budget optimum, and the multilevel telescoping of Section 4.1 provides no benefit here (its savings are realized only on noisier oracles, $\sigma > 0$). We note for completeness that estimating ρ_ℓ from real-engine traces requires the r replicate calls at a fixed prefix to be exchangeable: both serving stacks group queries into kernel-internal batches, so consecutive calls in the orchestrator's call order can share scheduling-induced state and exhibit short-range correlations along the r axis, and a naive split of the r-axis underestimates $\mathrm{Var}[Z_\ell]$; any such estimate must first apply an independent uniform permutation of the r-axis at each (trajectory, position) cell.

D.5. Empirical top-k support size

The query complexity bounds in the introduction are stated in terms of the support size K of the next-token distribution. When the API exposes a stated top-k parameter, the relevant K is the union of per-call top-k sets across the r repeated queries used by Algorithm 3 at a fixed prefix; nondeterminism in which tokens reach the threshold can make this union strictly larger than k. Figure 6 measures this union size on `Qwen3-0.6B` at the prefix of Appendix D.1, sweeping the stated top-$k \in \{5, 20, 100\}$ over r up to 248. Across all three settings the mean union remains within 7% of the stated k, so the practical K is comparable to the API-exposed value.

D.6. Calibration of σ^2 on a real LLM

The σ^2 in Definition 3 is the relative variance of a single noisy oracle output against the *true* prefix-conditional distribution p. On a real LLM we do not have p, so we substitute a high-quality empirical proxy and correct for the residual noise in the proxy.

Concretely, we reuse the $K_{\max} = 248$-repeat cache of Appendix D.5 on `Qwen3-0.6B` (bf16, SDPA), 100×500 (trajectory, position) cells. At each cell, we split the $K_{\max}$ iterations into two disjoint halves A, B of size $K_{\text{half}} = 124$. Let $\hat{p}_{r,A}$ be

the average of the first r iterations of A, and $\hat{p}_B$ the full average over B. Both are unbiased estimators of p, and they are independent because they use disjoint iterations. By Lemma 6 (the σ^2/r decay underlying Theorem 3), to leading order

$$\mathbb{E}\left[\chi^2(\hat{p}_{r,A} \,\|\, \hat{p}_B)\right] \;\approx\; \sigma^2 \left(\frac{1}{r} + \frac{1}{K_{\text{half}}}\right).$$

We measure the left-hand side empirically, average over cells, then divide by the right-hand side prefactor at each $r \in \{1, 2, 4, 8, 16, 32, 64, 124\}$ to obtain eight σ^2 estimates per top-k. If the noisy-oracle model holds, these eight estimates should coincide; the spread is therefore a self-consistency check. We observe agreement within $\pm 15\%$, indicating that σ^2 is an intrinsic property of the kernel rather than an artifact of the chosen averaging depth, and we report the across-r mean as the calibrated σ^2 in Figure 4. The K values in the same figure are the union saturation values from Figure 6 at $r = 248$.

E. Comparison between TV and KL

The comparison between TV and KL has a long history in probability, statistics, and information theory (Gibbs & Su, 2002; Sason & Verdú, 2016; Canonne, 2022). Here we sketch the aspects most relevant to our setting. We first recall several standard comparisons between total variation distance and KL divergence. A classical relation is Pinsker's inequality,

$$\mathrm{TV}\,(P, Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}\,(P\|Q)},$$

which shows that small KL divergence implies small TV distance. However, this relation is only one-sided and can be loose in several regimes. For example, two distributions may be arbitrarily close in TV while having infinite KL divergence if their supports do not match. Concretely, considering $P_\varepsilon = \mathrm{Bern}(\varepsilon), Q = \mathrm{Bern}(0)$, we have

$$\mathrm{TV}\,(P_\varepsilon, Q) = \varepsilon, \qquad \mathrm{KL}\,(P_\varepsilon\|Q) = +\infty.$$

This illustrates that KL is highly sensitive to support mismatch, whereas TV only charges the amount of probability mass on which the two distributions differ. Conversely, when two distributions are already very far apart, TV saturates at its maximum value 1, while KL, being unbounded, can still reflect different degrees of separation.

The two quantities also differ in their operational meanings. TV admits the optimal coupling characterization

$$\mathrm{TV}\,(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{P}_{(X,Y)\sim\gamma}[X \neq Y],$$

where $\Gamma(P, Q)$ denotes the set of all couplings of P and Q. Equivalently, there exists a maximal coupling such that

$$\mathbb{P}[X = Y] = 1 - \mathrm{TV}\,(P, Q)\,.$$

Thus, a statement such as $\mathrm{TV}\,(P, Q) = 0.01$ has a clear sample-level meaning: the two distributions can be coupled so that their samples are exactly identical with probability 0.99. This makes TV particularly natural for measuring observable output-level discrepancy between two generative distributions.

TV also has a direct decision-theoretic interpretation. Consider the binary Bayesian testing problem where one first samples a hidden label $H \in \{0, 1\}$ uniformly at random, and then observes $X \sim P$ if $H = 0$ and $X \sim Q$ if $H = 1$. The minimum possible probability of error over all decision rules is

$$P^\star_{\text{err}} = \frac{1 - \mathrm{TV}\,(P, Q)}{2}.$$

Thus, TV directly measures the distinguishability of two distributions. In other words, if two APIs are treated as black-box generators and we want to detect whether their output distributions differ, TV gives the fundamental statistical limit of distinguishability. A small TV implies that no black-box detector can reliably tell the two APIs apart from a single output sample, whereas a large TV implies that the two APIs are distinguishable with small Bayesian error.

Compared with TV, KL divergence and its variants are more widely used in empirical machine learning. For example, the cross-entropy loss used to train language models satisfies

$$\mathbb{E}_{X\sim P}[-\log Q_\theta(X)] = H(P) + \mathrm{KL}\,(P\|Q_\theta)\,,$$

and perplexity is an exponential transformation of this average log-loss. Thus, standard language-model training can be viewed as minimizing a KL-type objective.

This popularity is partly due to two practical advantages. First, KL admits an exact chain rule under autoregressive factorization:

$$\mathrm{KL}\left(P \| Q\right) = \sum_{t=1}^{n} \mathbb{E}_{X_{<t} \sim P}\left[\mathrm{KL}\left(P_t(\cdot \mid X_{<t}) \| Q_t(\cdot \mid X_{<t})\right)\right].$$

Thus, a sequence-level KL decomposes into token-level terms that are directly accessible from prefix probabilities. Second, KL-based objectives are smooth and naturally compatible with gradient-based training. In contrast, TV is bounded and operationally meaningful, but it lacks an analogous chain rule and is less convenient as a direct differentiable loss.