Large Language Models Are Bad Dice Players: LLMs Struggle to Generate Random Numbers from Statistical Distributions

Minda Zhao Yilun Du Mengyu Wang
Harvard University

Abstract

As large language models (LLMs) transition from chat interfaces to integral components of stochastic pipelines and systems approaching general intelligence, the ability to faithfully sample from specified probability distributions has become a functional requirement rather than a theoretical curiosity. We present the first large-scale, statistically powered audit of native probabilistic sampling in frontier LLMs, benchmarking 11 models across 15 distributions. To disentangle failure modes, we employ a dual-protocol design: Batch Generation, where a model produces N=1000 samples within one response, and Independent Requests, comprising N=1000 stateless calls. We observe a sharp protocol asymmetry: batch generation achieves only modest statistical validity, with a 7% median pass rate, while independent requests collapse almost entirely, with 10 of 11 models passing none of the distributions. Beyond this asymmetry, we reveal that sampling fidelity degrades monotonically with distributional complexity and aggravates as the sampling horizon N increases. Finally, we demonstrate how the propagation of these failures into downstream real-world application tasks introduces systematic biases: models fail to enforce uniform answer-position constraints in Multiple Choice Question generation and systematically violate demographic targets in attribute-constrained text-to-image prompt synthesis. These findings indicate that current LLMs lack a functional internal sampler, necessitating external tools for applications requiring statistical guarantees.

1 Introduction

As large language models (LLMs) transition from open-ended dialogue agents toward core components of complex application pipelines (Bubeck et al., 2023; Bommasani, 2021; Wang et al., 2024; Park et al., 2023; Li et al., 2023), their capacity for statistically faithful probabilistic sampling has emerged as a critical functional requirement (Gu et al., 2024). The prominence of LLMs in synthetic data generation (Li et al., 2023) has underscored the critical need for robust sampling mechanisms across diverse application scenarios (Shumailov et al., 2024). For instance, in educational material generation, the need for faithful sampling is particularly acute. Automatic question generation has emerged as a promising application of LLMs, with the potential to reduce instructor workload and enable personalized learning at scale (Kurdi et al., 2020). A critical requirement in Multiple Choice Question (MCQ) construction is that correct answers be uniformly distributed across positions (A, B, C, D) to prevent test-takers from exploiting positional patterns (Haladyna, 2004). Yet prior work has shown that LLMs exhibit strong positional preferences when selecting among options (Zheng et al., 2023; Wang et al., 2023). Whether analogous biases emerge during generation, when models must produce MCQs adhering to uniformity constraints, remains unexplored. Similarly, in text-to-image generation pipelines, LLMs are increasingly employed to automatically generate diverse prompt sets (Hao et al., 2023b; Rosenman et al., 2024). When constructing synthetic image datasets, practitioners often require prompts with controlled attribute distributions, such as demographic balance across gender and ethnicity, to ensure representational fairness in downstream applications (Sahili et al., 2024). The effectiveness of this approach hinges entirely on the LLM's ability to faithfully sample from specified distributions; if native sampling is biased, the resulting prompts will systematically deviate from target specifications. Currently, to ensure statistical rigor, the mainstream practice involves prompting LLMs to generate Python code that calls external numerical li-

Code and data are available at https://github.com/Mininda/LLM_Bad_Dice_Player.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22942–22959

braries such as `numpy.random` and this reliance on code-based workarounds is not incidental but systematic (Gao et al., 2023; Chen et al., 2023; Schick et al., 2023). However, the ability to internalize and simulate world dynamics is increasingly viewed as a prerequisite for general intelligence (LeCun, 2022). Just as the community has pursued making LLMs solve mathematical problems without external calculators (Wei et al., 2022; Lewkowycz et al., 2022; Shao et al., 2024), we ask whether models can generate samples from specified distributions without relying on external libraries. If a model must rely on an external calculator to generate even basic distributions, it suggests the model has learned to produce linguistic descriptions of randomness without acquiring the underlying functional competence (Mahowald et al., 2024).

Recent studies have begun to examine the native sampling capabilities of LLMs, yielding valuable but fragmented insights. Hopkins et al. (2023) identified systematic biases toward "favorite" numbers in random integer generation, while Xiao et al. (2025) revealed persistent deviations from target probabilities in coin-flip tasks. As the most extensive empirical study to date, Gu et al. (2024) evaluated five probability distributions within behavioral simulation contexts. However, their study is constrained by a sample size of $N = 100$ insufficient for reliable convergence assessment, and an experimental scope limited to five simple distributions. Moreover, their single-prompt protocol generates all samples in one response rather than through independent calls, this design cannot determine whether models possess genuine independent sampling capabilities.

To address these gaps, we present the first large-scale, systematic evaluation of native probabilistic sampling capabilities in frontier LLMs. Our benchmark characterizes the stochastic fidelity of 11 state-of-the-art models across a taxonomy of 15 probability distributions. Distinguishing our work from prior small-scale studies, we evaluate each configuration at a high-resolution sample size of N=1000, enabling a statistically powered assessment of distributional convergence. Central to our methodology is a dual-protocol experimental design intended to disentangle distinct failure modes: (1) Batch Generation, where the model generates a sequence of samples within a single context window, and (2) Independent Requests, where each sample is produced via an independent call. Beyond abstract distributional benchmarks,

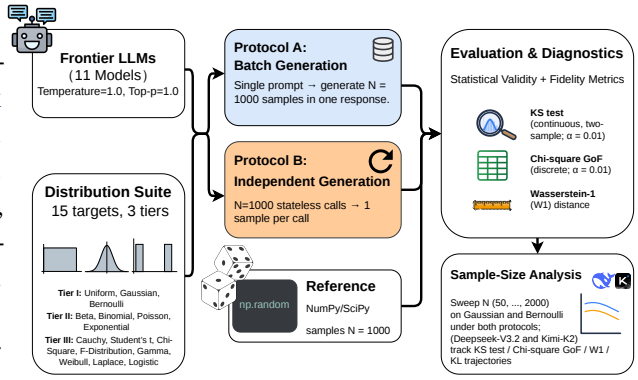

Figure 1: Overview of the Evaluation Pipeline. We systematically benchmark 11 frontier LLMs across 15 probability distributions spanning three complexity tiers. The evaluation employs a dual-protocol design to disentangle failure modes: *Protocol A (Batch)* produces samples sequentially within a single response, while *Protocol B (Independent)* produces samples via stateless single-sample calls. Distributional fidelity is rigorously quantified using statistical validity tests (KS, χ^2) and geometric metrics ($\mathcal{W}_1$) against high-precision `numpy` or `scipy` reference samples.

we provide the first systematic evidence that native sampling failures carry downstream consequences: in MCQ generation, models exhibit severe positional bias despite explicit uniformity instructions; in attribute-constrained prompt synthesis, demographic specifications are systematically violated. Our contributions are summarized as follows: (1) We demonstrate that current LLMs lack a functional internal mechanism for probabilistic sampling. (2) We reveal that sampling performance is bounded by distributional complexity. (3) We identify that increasing the sampling budget (N) degrades distributional adherence.

2 Related Work

Sampling in LLM Applications. LLMs now serve as core components in applications demanding statistical fidelity. In agent-based systems, Park et al. (2023) demonstrated that LLM-powered agents can simulate believable human behaviors in interactive environments, where behavioral diversity is essential for realistic social dynamics. Similarly, Hao et al. (2023a) showed that LLMs can function as world simulators, predicting environment state transitions to enable multi-step planning. These paradigms rely on the model's ability to faithfully sample from complex probability spaces, as precise stochasticity is essential

for maintaining behavioral diversity and modeling the inherent uncertainty of environment transitions. Beyond agents, a broad class of generative applications requires LLMs to produce outputs conforming to explicit distributional constraints. In synthetic data generation, researchers rely on LLMs to create diverse training sets (Li et al., 2023; Huang et al., 2025), yet biased sampling produces coverage gaps that propagate into downstream model failures (Shumailov et al., 2024). In educational material generation, LLMs are deployed to automatically construct assessments, exercises, and personalized learning materials at scale (Kurdi et al., 2020; Kasneci et al., 2023; Yan et al., 2024); applications requiring randomized test construction depend critically on the model's ability to honor uniformity constraints (Haladyna, 2004). In text-to-image pipelines, LLMs increasingly serve as prompt generators and parsers for diffusion models (Hao et al., 2023b; Qin et al., 2024), producing prompt sets with controlled attribute distributions for dataset construction (Rosenman et al., 2024); demographic balance requirements for fairness (Sahili et al., 2024) hinge entirely on faithful sampling from target specifications. While prior work has documented discriminative biases, positional preferences when LLMs select among options (Zheng et al., 2023; Wang et al., 2023), whether analogous biases emerge during generation, when models must produce content adhering to explicit distributional constraints, remains unexplored. Collectively, these applications demonstrate that native sampling fidelity is not a peripheral concern but a foundational prerequisite; its limitations directly dictate the statistical integrity of downstream generative systems, necessitating a rigorous and systematic audit.

Empirical Studies of LLM Sampling. A few recent studies have begun to provide preliminary evaluations of the randomness of LLM-generated outputs. Researchers have tested simple scenarios like prompting an LLM to generate uniform random bits or numbers, only to find significant deviations from true randomness (Hopkins et al., 2023). For instance, models often exhibit a favorite outcome instead of a uniform spread (Hopkins et al., 2023). Similarly, Xiao et al. (2025) demonstrates a knowledge–sampling gap in Bernoulli (coin-flip) tasks: across multiple frontier LLMs (e.g., Llama-3.1, GPT-4.1-nano, DeepSeekV3, Qwen-2.5), direct sampling from (0,1) remains systematically biased and highly sensitive to prompt phrasing, even when the target probability is explicitly specified. Despite focusing only on these basic Uniform and Bernoulli cases, prior work already reveals that LLMs struggle to generate target distributions. These limitations underscore the lack of a large-scale, statistically rigorous benchmark capable of verifying the native sampling mechanisms of foundation models across diverse distribution types. The most comprehensive effort to date, Gu et al. (2024) provided a more systematic evaluation by testing five distinct probability distributions within behavioral simulation contexts. Interestingly, their findings stand in tension with earlier reports of failure, ostensibly suggesting that frontier models can approximate target distributions. However, their analysis is constrained by a limited sample size (N=100) insufficient for assessing asymptotic convergence and an experimental scope restricted to only five elementary distributions. Moreover, their reliance on batch-generation fails to disentangle sequential inter-dependencies from native sampling, leaving the question of genuine independent stochasticity unanswered. Our work addresses these gaps through the first large-scale evaluation (N=1000) across 15 distributions and 11 models, employing dual protocols to disentangle batch and independent sampling failures.

3 Methodology

3.1 Problem Formulation: The Context–Fidelity Dilemma

We evaluate whether an LLM can faithfully sample from a user-specified 1D target distribution $\mathcal{P}$ over $\mathbb{R}$. Given a sampling budget N, the model returns samples $S_N = \{x_1, \ldots, x_N\}$, inducing an empirical measure $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ where δ_x is the Dirac measure . We measure fidelity by the Wasserstein-1 distance $\mathcal{W}_1(\hat{\mu}_N, \mu_\mathcal{P})$. For measures on $\mathbb{R}$ with a finite first moment, $\mathcal{W}_1$ admits the CDF form (Vallender, 1974):

$$\mathcal{W}_1(\hat{\mu}_N, \mu_\mathcal{P}) = \int_{-\infty}^{\infty} |F_{\hat{\mu}_N}(x) - F_\mathcal{P}(x)|\, dx, \tag{1}$$

where $F_{\hat{\mu}_N}$ is the empirical CDF induced by S_N and $F_\mathcal{P}$ is the target CDF. For an ideal i.i.d. sampler from $\mathcal{P}$, standard concentration results imply the expected error decreases with N at the canonical $\mathcal{O}(N^{-1/2})$ rate (Massart, 1990; Fournier and Guillin, 2015). Our experiments show that

LLMs systematically deviate from this baseline in a protocol-dependent way.

Let $\mathcal{E}(N) = \mathbb{E}[\mathcal{W}_1(\hat{\mu}_N, \mu_{\mathcal{P}})]$ denote the expected fidelity error.

Regime I: Independent Requests (stationary induced distribution). Under stateless calls conditioned on a fixed prompt and decoding configuration θ, outputs are modeled as conditionally i.i.d. draws from a stationary induced distribution $\mathcal{Q}_\theta$. Defining the intrinsic mismatch floor as $\Delta_{\text{ind}} := \mathcal{W}_1(\mu_{\mathcal{Q}_\theta}, \mu_{\mathcal{P}})$, the expected error stabilizes at:

$$\mathcal{E}(N) = \Delta_{\text{ind}} + \mathcal{O}(N^{-1/2}) \quad \text{as } N \to \infty. \tag{2}$$

The $\mathcal{O}(N^{-1/2})$ term represents sampling noise that vanishes asymptotically, leaving the irreducible bias Δ_{ind}.

Regime II: Batch Generation (Correction vs. Drift). In a single-response sequence, each x_i is drawn from a history-dependent conditional $\mathcal{Q}_\theta(\cdot \mid x_{<i})$. History dependence enables early self-correction, yet long-horizon autoregression risks accumulating deviation (Bengio et al., 2015; Ranzato et al., 2016). We capture this non-monotonicity through a diagnostic decomposition relative to a baseline horizon N_0, which denotes the smallest evaluated horizon:

$$\begin{aligned} \text{Corr}(N) &:= \max_{n \le N} \left(\mathcal{E}(N_0) - \mathcal{E}(n)\right), \\ \text{Drift}(N) &:= \mathcal{E}(N) - \min_{n \le N} \mathcal{E}(n). \end{aligned} \tag{3}$$

This formulation yields the identity:

$$\mathcal{E}(N) = \mathcal{E}(N_0) - \underbrace{\text{Corr}(N)}_{\text{Correction Gain}} + \underbrace{\text{Drift}(N)}_{\text{Exposure Bias}}. \tag{4}$$

The Context–Fidelity Dilemma arises because larger contexts can increase correction, yet beyond a critical horizon, the incremental increase in drift outweighs the incremental correction, causing net fidelity to degrade.

3.2 Metrics

Wasserstein-1 Distance ($\mathcal{W}_1$). Serving as our primary proxy for error $\mathcal{E}$, $\mathcal{W}_1$ measures the geometric cost to transport the generated mass to the target distribution. As established in Eq. (1), we compute the L_1 distance between CDFs.

KL Divergence. We measure information loss using a histogram-based approximation to $D_{\text{KL}}(\hat{p} \,\|\, p)$, where $\hat{p}$ denotes the empirical distribution of model samples and p denotes the target reference distribution.

Statistical Validity Tests. To enforce a conventional binary diagnostic criterion for validity ($p > \alpha$ with α=0.01), we apply distinct tests based on the support type: For continuous distributions, we employ the two-sample Kolmogorov–Smirnov (KS) test, comparing the empirical CDF of LLM-generated samples against that of high-precision reference samples ($N_{ref} = 1000$) drawn from the target distribution. For discrete distributions, we employ the Chi-square goodness-of-fit test (χ^2), comparing observed outcome counts against expected theoretical counts. We reject the null hypothesis (that the LLM and reference samples originate from the same underlying distribution) if the p-value falls below $\alpha = 0.01$.

4 Experiment

4.1 Models Under Evaluation

We benchmark eleven frontier language models representing diverse access paradigms and architectural families to ensure comprehensive coverage of the current LLM landscape. Our selection includes `GPT-5.2`, `GPT-4o` (Hurst et al., 2024), `GPT-OSS-120B` (Agarwal et al., 2025) (OpenAI), `Gemini-3-pro` (Comanici et al., 2025), `Gemma-3-27B` (Team et al., 2025a) (Google), `DeepSeek-V3.2` (Liu et al., 2024) (DeepSeek), `Kimi-K2` (Team et al., 2025b) (Moonshot), `Qwen3-32B` (Yang et al., 2025) (Alibaba), `Mistral-Small-3.2-24B` (Jiang et al., 2023) (Mistral AI), `Llama-3.3-70B` and `Llama-4-Scout-17B` (Grattafiori et al., 2024) (Meta). We use the standard default decoding settings ($T = 1.0$, `top-p` $= 1.0$), which are the most commonly used and provide a practical balance between randomness and output stability, making them a natural choice for evaluating native sampling ability. We further report decoding-parameter ablations in Appendix A. For downstream application experiments (MCQ generation and text-to-image prompt generation), we select six representative models: `GPT-4o`, `DeepSeek-V3.2`, `Qwen3-32B`, `Llama-3.3-70B`, `Llama-4-Scout`, and `GPT-OSS-120B`.

Tier	Distribution	Parameters	Diagnostic Target
I	Uniform	$a = 0,\ b = 1$	Range uniformity
	Gaussian	$\mu = 0,\ \sigma = 1$	Central tendency
	Bernoulli	$p = 0.7$	Binary asymmetry
II	Beta	$\alpha = 2,\ \beta = 2$	Bounded support $[0, 1]$
	Binomial	$n = 10,\ p = 0.5$	Discrete counting
	Poisson	$\lambda = 5$	Event rate modeling
	Exponential	$\lambda = 1$	Positive-only support
III	Cauchy	$x_0 = 0,\ \gamma = 1$	Undefined moments
	Student's t	$\nu = 3$	Fat tails
	Chi-Square	$\nu = 5$	Sum-of-squares
	F-Distribution	$d_1 = 5,\ d_2 = 10$	Ratio complexity
	Gamma	$\alpha = 2,\ \beta = 2$	Shape-scale
	Weibull	$k = 1.5,\ \lambda = 1$	Reliability modeling
	Laplace	$\mu = 0,\ b = 1$	Sharp peak
	Logistic	$\mu = 0,\ s = 1$	Sigmoid symmetry

Table 1: Distribution benchmark suite with parameters and diagnostic targets.

4.2 Distribution Sampling Evaluation

Distribution Taxonomy. To evaluate stochastic sampling behavior, we benchmark models on 15 probability distributions organized into three tiers based on entropy characteristics, support constraints, and tail behavior (Murphy, 2012; Gelman et al., 1995) (Table 1). Tier I includes canonical distributions such as Gaussian and Uniform, serving as standard building blocks in probabilistic modeling. Tier II covers distributions with bounded supports or discrete domains (e.g., Beta, Poisson), assessing adherence to strict validity constraints. Tier III comprises heavy-tailed or multi-parameter distributions (e.g., Student's t, Gamma), stress-testing robustness and tail-sensitive behavior beyond low-order moments. We further report a bivariate Gaussian experiment as a multivariate extension in Appendix B.

Sampling Protocols. We employ two complementary protocols to disentangle distinct failure mechanisms.

Protocol A: Batch Generation. Models receive a single prompt requesting N=1000 samples from the target distribution, generating all values within one response:

> *"You are a random number generator. Your task is to generate exactly {N} independent samples from a [Distribution] distribution with parameters [params]."*

This protocol forces the model to condition on its generated history, probing the cumulative effect of extended context on distributional fidelity.

Protocol B: Independent Requests. We issue N=1000 stateless calls, each requesting exactly one sample. The prompt is reduced to:

> *"Generate exactly ONE random number from a [Distribution] distribution with parameters [params]. Output ONLY the number."*

Each call is independent with no shared context, isolating the model's intrinsic priors without contextual interference.

Statistical Testing and Metrics. For continuous distributions, we apply the two-sample Kolmogorov-Smirnov test comparing N=1000 LLM-generated samples against N=1000 reference samples from `numpy.random` and `scipy.stats`; for discrete distributions, we use Chi-square goodness-of-fit against theoretical PMFs. All tests use $\alpha = 0.01$ following Gu et al. (2024), as our large sample size (N=1000) increases statistical power, requiring a stricter significance threshold to prevent the over-interpretation of minor deviations as substantive failures. A sensitivity analysis with alternative significance thresholds is provided in Appendix D. We additionally report Wasserstein-1 distance ($\mathcal{W}_1$) and KL divergence for fine-grained fidelity quantification.

Sample-Size Scaling Analysis. To characterize convergence trajectories and identify collapse thresholds, we sweep sample sizes $N \in \{50, 100, 200, 300, \ldots, 1000, 1500, 2000\}$ for Gaussian and Bernoulli sampling with DeepSeek-V3.2 and Kimi-K2 under both batch and independent protocols. At each checkpoint, we report $\mathcal{W}_1$, KL divergence, and the corresponding goodness-of-fit statistic: Kolmogorov-Smirnov (KS) for Gaussian and χ^2 for Bernoulli.

4.3 Downstream Applications

MCQ Generation. To examine whether sampling deficiencies propagate to structured generation, we design an MCQ benchmark requiring models to produce N=1000 medical multiple-choice questions via independent calls. Crucially, prompts explicitly instruct that the position of the correct answer should be randomly and uniformly distributed among A, B, C, D to ensure no positional bias. We extract designated correct answers' position from each generated question and perform χ^2 goodness-of-fit tests against the uniform target (25% per option). This directly tests whether models can internalize uniformity constraints during content creation.

Attribute-Constrained Prompt Generation. We further stress-test native sampling in a semantically grounded setting where distributional constraints are entangled with natural language generation. Each model generates $N{=}1000$ text-to-image prompts via independent calls, each describing a person wearing a coat. Four attributes must independently conform to prescribed target distributions: Gender (Male 49.5%, Female 50.5%) and Race/Ethnicity (White 57.5%, Hispanic 20.0%, Black 12.6%, Asian 6.5%, Other 3.4%) derived from U.S. Census Bureau (2024); Height following $\mathcal{N}(169, 10^2)$ cm; and Coat Color uniformly distributed over seven categories. We apply χ^2 tests for categorical attributes and KS tests for height. This task evaluates whether LLMs can faithfully sample from explicit distributional specifications when probability constraints must be realized through semantically coherent text.

Dist.	GPT-5.2	Gemini-3	GPT-4o	DeepSeek	Qwen3	Gemma-3	Mistral-3.2	Kimi-K2	Llama-3.3	Llama-4	GPT-OSS
Discrete Distributions											
Bernoulli	0.08	0.04	**4e-05***	0.02	0.12	0.06	0.07	**0.03***	0.16	0.13	**0.02***
Binomial	0.59	0.19	0.26	0.20	0.80	1.1	0.53	0.81	0.32	0.56	0.58
Poisson	0.64	0.32	0.26	0.21	1.5	1.0	0.95	0.62	0.60	0.62	0.71
Continuous Distributions											
Uniform	**0.02***	**0.01***	**0.02***	**9e-03***	0.10	**0.03***	0.15	0.03	0.07	0.19	**0.03***
Gaussian	**0.13***	0.15	**0.10***	0.43	0.21	0.15	0.31	0.17	0.36	0.26	0.23
Beta	0.08	0.06	0.10	0.06	0.19	0.06	0.06	**0.03***	0.08	0.10	0.08
Exp	0.43	0.11	0.24	0.30	0.21	0.38	0.86	0.52	0.25	0.35	0.42
Cauchy	5.4	5.0	**3.3***	5.9	5.4	6.6	6.1	5.7	6.4	6.0	6.0
t	0.75	0.46	**0.13***	0.48	0.72	0.48	0.43	**0.26***	0.37	0.50	0.52
χ^2	0.75	1.0	0.93	1.2	1.5	1.6	4.5	1.8	0.91	**0.66***	0.59
F	0.43	0.20	0.45	0.14	0.40	0.98	0.71	0.79	0.42	0.50	0.90
Gamma	0.33	1.0	0.78	0.84	1.3	2.0	1.5	1.5	2.4	1.8	1.7
Weibull	0.29	0.13	0.32	0.36	0.34	0.13	0.37	0.51	0.67	0.33	0.33
Laplace	0.28	**0.15***	**0.32***	0.35	1.2	0.35	0.69	0.44	0.46	0.50	0.74
Logistic	0.99	0.37	0.73	0.54	2.1	0.40	0.51	0.39	1.7	0.92	0.52
Pass Rate	**13%**	**13%**	**40%**	**7%**	0%	**7%**	0%	**20%**	0%	**7%**	**13%**

Table 2: Wasserstein Distance $\mathcal{W}_1$ (Batch Generation). Lower $\mathcal{W}_1$ indicates better distributional fit. * denotes passing the statistical test ($p > 0.01$): χ^2 GoF for discrete, two-sample KS for continuous.

5 Results

5.1 Distribution Sampling

Protocol-Dependent Sampling Fidelity. Table 2 shows that batch generation achieves modest statistical validity: the leading model passes 40% of distributions, while the median pass rate is 7%. In stark contrast, Table 3 shows a near-complete failure under independent sampling, with 10 of the 11 models failing to pass any distribution. This protocol asymmetry is not attributable to distributional difficulty: examining Uniform (the simplest benchmark), Wasserstein distances amplify from $\mathcal{W}_1{\approx}0.01$ (batch) to $\mathcal{W}_1{\approx}0.15$ (independent) across models. The stark contrast in performance suggests that valid sampling depends critically on long-context dependencies, rather than being reliably supported by isolated, stateless sampling.

Complexity Stratification. We further stratify distributions by complexity tier to examine how sampling fidelity varies with increasing distribution complexity. Figure 2 visualizes this dual trend: panel (a) shows pass rates declining monotonically across tiers for most models, with GPT-4o exhibiting steepest degradation from perfect Tier I performance; panel (b) demonstrates the inverse relationship, $\mathcal{W}_1$ distances rise systematically with tier complexity, diverging from $\sim$0.1 (Tier I) to $\sim$1.5 (Tier III) across models. This coupled pattern, with declining validity alongside escalating distributional distance, shows that increasing structural constraints are associated with progressively worse performance across distribution tiers.

Sample-Size Trajectories. To examine how sampling budget affects distributional adherence, we analyze Gaussian and Bernoulli generation trajectories for $N \in \{50, \ldots, 2000\}$ (Figure 3 and 4). The results reveal a distinct inverse scaling trend. While initial fluctuations in KL and $\mathcal{W}_1$ are attributable to finite-sample variance, the long-run behavior contradicts standard convergence expectations. Batch generation exhibits pronounced degradation: as N exceeds 1000, $\mathcal{W}_1$ distances rise steadily, coinciding with KS p-values collapsing below the significance threshold ($\alpha = 0.01$). Although Independent requests fail the KS test across all N, they display a parallel drift, with $\mathcal{W}_1$ increasing gradually. These trajectories confirm that for current LLMs, larger sample sizes reveal the statistically significant discrepancy that is invisible at small N.

5.2 Downstream Applications

MCQ Generation: Positional Bias Persists Despite Explicit Instructions. Table 4 quantifies generative positional bias in MCQ construction. Despite explicit prompts requiring uniform distribution of correct answers across A/B/C/D positions, all six models exhibit severe and statistically significant bias ($p < 0.001$). GPT-OSS-120B shows the most extreme skew, placing 54.6% of correct answers at position C and only 4.5% at position A. GPT-4o favors position B (46.8%) while nearly ignoring D (5.5%). Notably, no model approaches

Dist.	GPT-5.2	Gemini-3	GPT-4o	DeepSeek	Qwen3	Gemma-3	Mistral	Kimi-K2	Llama-3.3	Llama-4	GPT-OSS
Discrete Distributions											
Bernoulli	0.32	0.32	0.31	0.12	0.29	0.32	0.27	0.18	0.32	**0.02***	0.32
Binomial	1.2	0.80	0.83	1.2	5.0	1.4	1.5	0.87	1.4	1.2	1.0
Poisson	1.0	0.98	0.52	0.99	1.7	2.5	0.82	0.71	2.3	1.1	1.5
Continuous Distributions											
Uniform	0.15	0.20	0.16	0.15	0.51	0.17	0.17	0.17	0.34	0.28	0.16
Gaussian	0.57	0.70	0.27	0.49	0.82	0.72	0.40	0.55	0.44	0.83	0.54
Beta	0.13	0.14	0.08	0.11	1.5	0.15	0.10	0.16	0.17	0.11	0.13
Exp	0.61	0.50	0.27	0.19	0.71	0.49	0.61	0.40	0.52	0.44	0.45
Cauchy	3.1	3.2	2.8	8.3	3.7	3.6	3.0	3.0	3.7	3.8	2.7
t	0.59	0.56	0.40	0.95	1.1	1.4	0.60	0.39	1.1	1.2	0.68
χ^2	1.7	1.7	1.1	11.6	2.5	2.9	1.6	1.1	2.1	3.1	1.8
F	0.54	0.56	0.56	3.7	3.5	1.2	0.49	0.32	0.65	0.31	0.57
Gamma	1.1	1.7	1.2	4.8	2.4	1.8	1.1	1.5	1.7	2.1	1.4
Weibull	0.35	0.40	0.30	2.3	0.75	0.47	0.50	0.26	0.40	0.31	0.26
Laplace	0.61	0.71	0.36	0.95	1.0	0.64	0.65	0.49	1.1	1.0	0.51
Logistic	0.88	0.80	0.66	0.75	1.3	1.2	0.91	0.90	1.3	1.4	0.60
Pass Rate	0%	0%	0%	0%	0%	0%	0%	0%	0%	**7%**	0%

Table 3: Wasserstein Distance $\mathcal{W}_1$ (Independent Requests). Lower $\mathcal{W}_1$ indicates better distributional fit. * denotes passing the statistical test ($p > 0.01$): χ^2 GoF for discrete, two-sample KS for continuous.

Model	A (%)	B (%)	C (%)	D (%)	χ^2	p
GPT-4o	12.6	46.8	35.1	5.5	444.5	<.001
Llama-3.3-70B	17.2	32.9	42.9	7.0	307.1	<.001
DeepSeek-V3.2	16.9	28.2	36.8	18.1	105.1	<.001
Qwen3-32B	21.8	35.6	31.5	11.1	143.2	<.001
Llama-4-Scout	28.6	42.3	22.7	6.4	265.4	<.001
GPT-OSS-120B	4.5	27.7	54.6	13.2	577.2	<.001
Uniform	25.0	25.0	25.0	25.0	–	–

Table 4: MCQ Answer Distribution Bias (English, Temperature=1.0, N=1000). Target: Uniform 25% per option. All models show significant bias ($p < 0.001$).

the uniform 25% target for any position. These results demonstrate that sampling deficiencies are not confined to abstract numerical generation but propagate directly into structured content creation, fundamentally compromising the reliability of LLM-generated evaluation materials.

Attribute-Constrained Prompt Generation: Systematic Distributional Violations. Table 5 reveals pervasive failures when models must translate explicit distributional specifications into semantically coherent text. For Gender, models exhibit opposing biases: GPT-4o overrepresents males (75.0% vs. target 49.5%), while Llama-4 drastically overrepresents females (97.2%). For Race/Ethnicity, models systematically over-sample certain groups (GPT-4o: 33.5% Asian vs. target 6.5%) while severely under-representing others (GPT-4o: 0% Hispanic vs. target 20.0%; Other category: 0% across four of six models vs. target

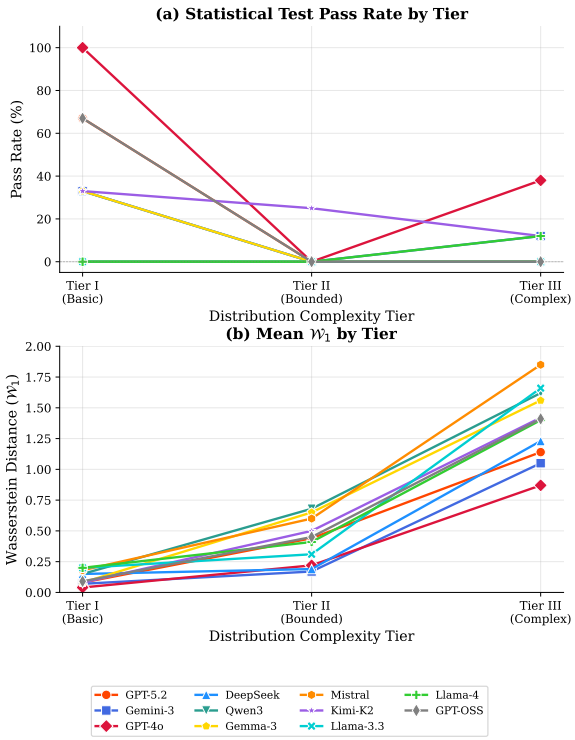

Figure 2: Distribution Complexity vs. Sampling Fidelity. (a) Statistical test pass rate decreases as distribution complexity increases from Tier I (Fundamental Priors) to Tier III (Heavy-Tailed & Complex). (b) Mean Wasserstein distance $\mathcal{W}_1$ increases with complexity, indicating poorer distributional fit.

3.4%). The Height distribution reveals variance collapse: all models produce $\sigma \approx 1$–6 cm versus the target $\sigma = 10$ cm, with KS statistics exceeding 0.37 across all models. For Coat Color, models collapse onto modal preferences, with Llama-3.3 generating 96% green coats and GPT-OSS favoring red (54%), completely ignoring the uniform specification. These failures persist despite prompts containing precise numerical targets, confirming that LLMs cannot internalize distributional constraints when sampling must occur through natural language generation rather than raw numerical output.

6 Discussion

LLMs Lack a Functional Internal Sampler. Our results provide compelling evidence that current LLMs do not possess a genuine internal mechanism for probabilistic sampling. The most striking finding is the near-total failure under independent sampling: 10 of 11 models achieve exactly 0% pass rate when generating samples without shared context. This stands in sharp contrast to batch gen-

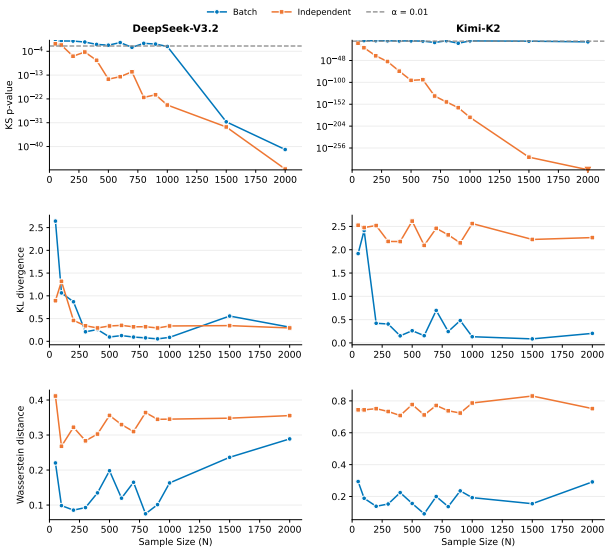

Figure 3: Effect of sample size (N) on the Gaussian sampling quality of DeepSeek-V3.2 and Kimi-K2. The dashed line indicates the Kolmogorov-Smirnov (KS) test significance threshold at $\alpha = 0.01$. Downward triangles in the Kimi-K2 KS p-value panel indicate p-values below the double-precision normal limit (2.23×10^{-308}.)

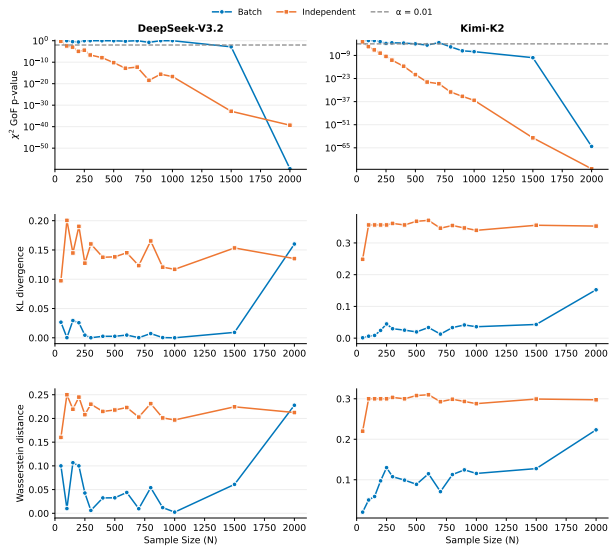

Figure 4: Effect of sample size (N) on the Bernoulli sampling quality of DeepSeek-V3.2 and Kimi-K2. The dashed line indicates the χ^2 goodness-of-fit test significance threshold at $\alpha = 0.01$.

eration, where models achieve modest pass rates through within-context self-correction. The implication is clear: what appears to be sampling capability in batch mode is in fact an emergent property of autoregressive conditioning, not an internalized understanding of probability distributions. When this contextual scaffolding is removed, models default to systematic internal biases that produce statistically invalid outputs. The apparent stochasticity of LLM outputs is therefore not grounded in distributional competence.

Complexity Amplifies Failure. Figure 2 reveals a consistent relationship between distribution complexity and sampling failure. Pass rates decline from Tier I to Tier III, while mean $\mathcal{W}_1$ increases correspondingly. Heavy-tailed distributions such as Cauchy and Chi-Square prove particularly challenging, with no model passing the statistical tests. This pattern suggests that LLMs can only approximate distributional forms that are well-represented in their training corpora. When confronted with mathematically complex distributions requiring precise handling of bounded supports, undefined moments, or multi-parameter dependencies, models fail to generalize beyond superficial pattern matching. The fidelity gap between simple and complex distributions underscores a fundamental limitation: LLMs learn to mimic the surface statistics of familiar distributions without acquiring the underlying mathematical structure.

Inverse Scaling Under Increasing Sampling Budget. Increasing sample size should improve distributional convergence. Instead, contrary to asymptotic convergence, distributional fidelity degrades as the requested horizon N grows. Figure 3 and 4 reveal a consistent inverse-scaling signature across diagnostics. In batch generation, $\mathcal{W}_1$ exhibits a clear regime shift: after an early improvement at short horizons, it turns upward and increases with N, consistent with length-amplified degradation in long sequences. Crucially, this is not a batch-only artifact. Under independent requests, the model is already invalid at small N (KS p-values below threshold), yet $\mathcal{W}_1$ still drifts upward with N, indicating that larger budgets expose progressively larger geometric mismatch even without shared context. As N increases, the accumulated discrepancy becomes statistically undeniable, driving KS p-values to vanishing levels. These findings reveal that expanding the sample budget unmasks fundamental distributional mismatches that remain statistically latent in smaller samples, particularly within batch generation regimes.

7 Limitations

Our findings are empirical rather than theoretical: they demonstrate that current frontier LLMs lack re-

	GPT-4o	DeepSeek	Qwen3	Llama-3.3	Llama-4	GPT-OSS	Target
Task: Gender							
Male (%)	75.0	60.4	29.3	31.7	2.8	21.3	49.5
Female (%)	25.0	39.6	70.7	68.3	97.2	78.7	50.5
χ^2	260.3	47.6	163.1	126.6	872.1	317.9	–
Task: Race/Ethnicity							
White (%)	41.8	54.5	57.2	70.5	51.3	45.0	57.5
Hispanic (%)	0.0	6.0	27.1	6.5	42.9	10.3	20.0
Black (%)	24.7	10.7	10.6	6.0	1.1	15.6	12.6
Asian (%)	33.5	28.5	4.9	17.0	4.7	29.1	6.5
Other (%)	0.0	0.3	0.2	0.0	0.0	0.0	3.4
Task: Height							
μ (cm)	172.7	171.3	171.9	172.5	172.5	173.0	169.0
σ (cm)	3.0	5.6	3.9	0.9	1.5	3.9	10.0
KS stat	0.51	0.37	0.39	0.61	0.53	0.52	–
Task: Coat Color							
Black (%)	1	8	2	0	4	0	14.3
White (%)	3	15	2	0	8	0	14.3
Red (%)	28	23	39	0	1	54	14.3
Blue (%)	0	1	20	0	19	2	14.3
Green (%)	54	28	28	96	55	29	14.3
Yellow (%)	11	13	4	0	7	14	14.3
Brown (%)	4	11	5	2	7	1	14.3

Table 5: Distribution fidelity in text-to-image prompt generation (N=1000). Models are prompted to sample attributes according to U.S. Census targets (Gender, Race) and standard statistical forms (Height, Color). All deviations are significant ($p < 0.001$).

liable native sampling under standard decoding, but do not constitute an impossibility proof for future architectures or training paradigms. Although we benchmark 15 canonical distributions across multiple complexity tiers, all targets are explicitly specified; real-world stochastic processes may involve implicit or context-dependent distributions beyond our experimental scope. Finally, the downstream tasks are used as controlled tests with explicit distributional constraints, to show that sampling failures alone can induce generation-stage bias, rather than to provide a comprehensive fairness analysis.

8 Ethical Considerations

To our knowledge, this work is the first to systematically demonstrate how native sampling failures propagate into downstream bias during the generation process. In the MCQ experiment, models' inability to follow uniform positional constraints directly produces answer-position bias, compromising the fairness of LLM-generated evaluation materials. In the attribute-constrained generation experiment, models' failure to adhere to demographic distributions causes sampling-induced bias to be embedded directly into synthetic data. These findings carry significant implications for high-stakes applications. When LLMs are deployed for social simulation, synthetic data generation, or randomized decision-making, their outputs are often implicitly treated as valid probabilistic samples. Our results demonstrate that this assumption is fundamentally unwarranted. We urge the community to critically examine the potential consequences of sampling infidelity in application contexts where distributional accuracy is essential for fairness, validity, or safety.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28.

Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Preprint*, arXiv:2211.12588.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Nicolas Fournier and Arnaud Guillin. 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. *Preprint*, arXiv:2211.10435.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. 1995. *Bayesian data analysis*. Chapman and Hall/CRC.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jia Gu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024. Do llms play dice? exploring probability distribution sampling in large language models for behavioral simulation. *Preprint*, arXiv:2404.09043.

Thomas M Haladyna. 2004. *Developing and validating multiple-choice test items*. Routledge.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023a. Reasoning with language model is planning with world model. In *EMNLP*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023b. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939.

Aspen K Hopkins, Alex Renda, and Michael Carbin. 2023. Can llms generate random numbers? evaluating llm sampling in controlled domains. In *ICML 2023 workshop: sampling and optimization in discrete space*.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Jianfeng Gao, Chaowei Xiao, Lichao Sun, and Xiangliang Zhang. 2025. Datagen: Unified synthetic dataset generation via large language models. *Preprint*, arXiv:2406.18966.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International journal of artificial intelligence in education*, 30(1):121–204.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *Preprint*, arXiv:2310.07849.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Preprint*, arXiv:2301.06627.

Pascal Massart. 1990. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. 2024. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.

Shachar Rosenman, Vasudev Lal, and Phillip Howard. 2024. Neuroprompts: An adaptive framework to optimize prompts for text-to-image generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 159–167.

Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. 2024. Faircot: Enhancing fairness in text-to-image generation via chain of thought reasoning with multimodal large language models. *arXiv preprint arXiv:2406.09070*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The curse of recursion: Training on generated data makes models forget. *Preprint*, arXiv:2305.17493.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025b. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

U.S. Census Bureau. 2024. American community survey (acs). `https://www.census.gov/programs-surveys/acs`.

S. S. Vallender. 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Bingyi Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tim Z. Xiao, Johannes Zenn, Zhen Liu, Weiyang Liu, Robert Bamler, and Bernhard Schölkopf. 2025. Flipping against all odds: Reducing llm coin flip bias via verbalized rejection sampling. *Preprint*, arXiv:2506.09998.

Nan Xu and Xuezhe Ma. 2025. Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems. *Preprint*, arXiv:2410.14166.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason E Weston, and Jing Xu. 2025. Following length constraints in instructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24232–24243, Suzhou, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 4659–4675.

A Decoding Parameter Ablations

Table 6 and 7 present the results of our decoding ablations (GPT-4o and Gemma; all 15 distributions; Batch + Independent; $N{=}1000$) over temperature $\in \{0.2, 0.5, 1.0, 1.2\}$ and `top-p` $\in \{0.9, 0.95, 1.0\}$. We did not include `top_k` in the cross-model ablation because `top-k` is not exposed consistently across all APIs like GPT-4o. The overall conclusion remains unchanged: decoding sweeps do not consistently recover faithful native sampling, and the Independent protocol remains broadly poor across settings, which strengthens our main claim that LLMs lack a functional internal sampler.

B Multivariate Extension

We further extend our evaluation to a bivariate Gaussian setting. Specifically, we consider a zero-mean Gaussian with correlation $\rho = 0.5$:

$$(X, Y) \sim \mathcal{N}(\mathbf{0}, \Sigma), \qquad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

In addition to the two marginals, we evaluate the conditional distribution $Y \mid X > 0$. We test three models (GPT-4o, Kimi-K2, and Gemma-3-27B) under both batch and independent settings with

Distribution	GPT-4o				Gemma-3			
	$T = 0.2$	$T = 0.5$	$T = 1.0$	$T = 1.2$	$T = 0.2$	$T = 0.5$	$T = 1.0$	$T = 1.2$
Batch Mode								
Discrete Distributions								
Bernoulli	0.28	0.28	4e-05*	0.11	0.03	0.08	0.06	0.09
Binomial	1.3	1.3	0.26	0.44	0.17	1.3	1.1	1.3
Poisson	1.2	1.7	0.26	0.78	1.0	1.8	1.0	0.77
Continuous Distributions								
Uniform	0.04	0.05	0.02*	0.01*	0.03	0.07	0.03*	0.02*
Gaussian	0.16	0.23	0.10*	1.9*	0.16	0.15	0.15	0.10*
Beta	0.17	0.10	0.10	0.02*	0.06	0.06	0.06	0.05
Exponential	1.0	1.0	0.24	–	0.34	0.37	0.38	0.37
Cauchy	6.1	6.2	3.3*	4.3*	11.9	5.7	6.6	5.5
t	0.33	0.41	0.13*	–	0.38	5.8	0.48	0.32*
χ^2	1.6	2.4	0.93	1.2	11.7	0.79	1.6	0.57*
F	0.78	0.59	0.45	4.9	16.8	2.7	0.98	6.0
Gamma	1.8	2.1	0.78	1.2	1.2	6.7	2.0	3.2
Weibull	0.88	0.29	0.32	1.6	0.31	0.37	0.13	3.5
Laplace	0.26	0.97	0.32*	0.32	0.25	0.34	0.35	0.31
Logistic	1.4	1.2	0.73	–	6.5	0.70	0.40	3.9
Independent Mode								
Discrete Distributions								
Bernoulli	0.32	0.31	0.31	0.21	0.32	0.32	0.32	0.32
Binomial	1.1	0.84	0.83	0.71	1.5	1.4	1.4	0.75
Poisson	1.1	0.75	0.52	0.44	3.3	3.3	2.5	3.2
Continuous Distributions								
Uniform	0.18	0.17	0.16	0.16	0.21	0.17	0.17	0.14
Gaussian	0.60	0.44	0.27	0.25	0.81	0.81	0.72	0.75
Beta	0.12	0.10	0.08	0.08	0.21	0.21	0.15	0.13
Exponential	0.62	0.57	0.27	0.45	0.68	0.62	0.49	0.46
Cauchy	3.7	3.4	2.8	2.5*	3.8	3.7	3.6	3.7
t	0.99	0.53	0.40	0.59	1.3	1.3	1.4	1.1
χ^2	2.2	1.6	1.1	1.2	2.5	2.1	2.9	1.9
F	1.3	1.3	0.56	1.3	3.7	3.4	1.2	3.0
Gamma	2.0	2.0	1.2	1.8	2.1	2.0	1.8	1.8
Weibull	0.42	0.39	0.30	0.33	0.75	0.75	0.47	0.68
Laplace	0.97	0.60	0.36	0.25	1.1	1.0	0.64	0.86
Logistic	1.2	1.0	0.66	0.68	1.3	1.3	1.2	1.2

Table 6: **Temperature Ablation.** Each cell reports Wasserstein distance $\mathcal{W}_1$; * indicates passing the corresponding statistical test ($p > 0.01$). For $T = 1.2$, GPT-4o produced garbled and unparseable outputs for the Exponential, t-distribution, and Logistic settings in five repeated attempts in batch mode; these cases are therefore reported as "–".

N=1000. To assess multivariate fidelity, we measure: (i) marginal two-sample KS and $\mathcal{W}_1$ for X and Y; and (ii) conditional two-sample KS and $\mathcal{W}_1$ for $Y \mid X > 0$. As shown in Table 8, the multivariate extension reveals the same qualitative failure mode as our main experiments. In batch mode, all three models pass the KS test for the Y marginal, yet still fail to recover the conditional distribution $Y \mid X > 0$. Under independent requests, fidelity degrades further across both marginals and the conditional distribution.

C Detailed Sampling Fidelity Results

As shown in Section 5, Figure 2 illustrates the degradation of sampling fidelity as distributional complexity increases. Table 9 provides the comprehensive numerical breakdown of this trend, detailing Pass Rates and Wasserstein-1 ($\mathcal{W}_1$) distances for all 11 models across the three complexity tiers under both Batch and Independent protocols. The data confirms a monotonic degradation in performance as complexity increases.

D Significance-Threshold Sensitivity

To assess sensitivity to the choice of significance threshold, we additionally report pass rates under $\alpha = 0.05$ and $\alpha = 0.005$, alongside our main setting $\alpha = 0.01$. As shown in Table 10, the results under $\alpha = 0.005$ remain broadly similar to those under $\alpha = 0.01$, and the qualitative trends as well as our main conclusions remain unchanged. This indicates that our findings are not an artifact of a particular threshold choice.

E Distribution-Sampling Prompt Example

To ensure reproducibility, we provide a representative example of the original batch prompt used in our distribution-sampling experiments. The prompt below shows the exact instruction style used to ask the model to generate N samples from a target distribution. Since prior studies suggest that LLMs

Distribution	GPT-4o			Gemma-3		
	$p = 0.9$	$p = 0.95$	$p = 1.0$	$p = 0.9$	$p = 0.95$	$p = 1.0$
Batch Mode						
Discrete Distributions						
Bernoulli	0.10	6e-03*	4e-05*	0.12	0.19	0.06
Binomial	0.91	0.85	0.26	1.0	1.4	1.1
Poisson	0.68	0.38	0.26	0.90	1.7	1.0
Continuous Distributions						
Uniform	0.09	0.02*	0.02*	0.01*	0.10	0.03*
Gaussian	0.31	0.20	0.10*	0.18	0.14	0.15
Beta	0.17	0.13	0.10	0.05	0.03	0.06
Exponential	1.0	0.50	0.24	0.60	0.37	0.38
Cauchy	5.8	5.8	3.3*	5.6	5.6	6.6
t	0.82	0.44	0.13*	0.34	2.7	0.48
χ^2	1.9	3.1	0.93	0.97	0.64	1.6
F	0.47	0.31	0.45	0.47	0.30	0.98
Gamma	1.5	1.2	0.78	2.0	1.1	2.0
Weibull	0.46	0.25	0.32	0.29	0.31	0.13
Laplace	0.44	0.38	0.32*	0.32	4.0	0.35
Logistic	0.92	0.79	0.73	0.71	0.97	0.40
Independent Mode						
Discrete Distributions						
Bernoulli	0.31	0.26	0.31	0.32	0.32	0.32
Binomial	0.77	0.88	0.83	0.72	0.70	1.4
Poisson	0.71	0.58	0.52	3.3	3.2	2.5
Continuous Distributions						
Uniform	0.17	0.17	0.16	0.14	0.13	0.17
Gaussian	0.31	0.28	0.27	0.81	0.79	0.72
Beta	0.09	0.09	0.08	0.16	0.19	0.15
Exponential	0.55	0.54	0.27	0.52	0.47	0.49
Cauchy	3.1	2.9	2.8	3.7	3.7	3.6
t	0.55	0.60	0.40	1.3	1.2	1.4
χ^2	1.4	1.3	1.1	1.9	2.0	2.9
F	1.3	1.2	0.56	3.2	3.1	1.2
Gamma	1.9	1.9	1.2	2.0	1.9	1.8
Weibull	0.38	0.36	0.30	0.73	0.73	0.47
Laplace	0.36	0.38	0.36	0.91	0.87	0.64
Logistic	0.87	0.83	0.66	1.2	1.2	1.2

Table 7: **Top-p Ablation.** Each cell reports Wasserstein distance $\mathcal{W}_1$; * indicates passing the corresponding statistical test ($p > 0.01$).

Model	Mode	X	Y	$Y \mid X > 0$
GPT-4o	Batch	1.4486	0.0496*	0.2873
Kimi-K2	Batch	0.3115	0.2306*	0.5258
Gemma-3-27B	Batch	0.3395	0.1812*	0.3994
GPT-4o	Independent	0.4008	0.4480	0.2969
Kimi-K2	Independent	0.3354	0.6079	0.2934
Gemma-3-27B	Independent	0.9125	0.6835	0.7576

Table 8: **Bivariate Gaussian Extension ($\rho = 0.5$, N=1000).** Each cell reports Wasserstein distance $\mathcal{W}_1$. Columns X and Y denote marginal results, and $Y \mid X > 0$ denotes the conditional result. * indicates that the corresponding two-sample KS test passes ($p > 0.01$).

do not reliably follow exact counting or length constraints (Yuan et al., 2025; Xu and Ma, 2025), we incorporated explicit instruction-level requirements into the prompt to encourage adherence to the requested number of samples and the prescribed output format.

F Downstream Application Experimental Prompts

To ensure reproducibility, we provide the exact system instructions and user prompts employed in our downstream application experiments in Table 12 and Table 13.

G Fine-Grained Attribute Analysis

To isolate each model's capability in distribution-constrained generation, we conducted four independent attribute sampling experiments testing Gender, Race/Ethnicity, Height, and Coat Color separately. For each attribute, we prompted five state-of-the-art LLMs to generate $N = 1000$ independent samples following explicitly specified target distributions. Unlike the joint experiment where models must simultaneously control multiple attributes, these independent tests measure single-attribute adherence in isolation. The target distributions are identical to those used in the joint experiment. We provided explicit distribution constraints and sampling instructions in each prompt to ensure models were

Model	Batch Mode								Independent Mode							
	Tier I (3)		Tier II (4)		Tier III (8)		Overall (15)		Tier I (3)		Tier II (4)		Tier III (8)		Overall (15)	
	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$	Rate	$\mathcal{W}_1$
Proprietary Models																
GPT-5.2	**67%**	0.08	0%	0.44	0%	1.14	**13%**	0.74	0%	0.35	0%	0.76	0%	1.11	0%	0.86
Gemini-3	**33%**	0.07	0%	0.17	**12%**	1.05	**13%**	0.62	0%	0.41	0%	0.60	0%	1.20	0%	0.88
GPT-4o	**100%**	0.04	0%	0.22	**38%**	0.87	**40%**	0.53	0%	0.24	0%	0.43	0%	0.92	0%	0.65
Open-Weights Models																
DeepSeek-V3.2	**33%**	0.15	0%	0.19	0%	1.23	**7%**	0.74	0%	0.25	0%	0.62	0%	4.17	0%	2.44
Qwen3	0%	0.15	0%	0.68	0%	1.62	0%	1.08	0%	0.54	0%	2.24	0%	2.04	0%	1.79
Gemma-3	**33%**	0.08	0%	0.65	0%	1.56	**7%**	1.02	0%	0.40	0%	1.14	0%	1.66	0%	1.27
Mistral-3.2	0%	0.18	0%	0.60	0%	1.85	0%	1.18	0%	0.28	0%	0.76	0%	1.10	0%	0.85
Kimi-K2	**33%**	0.08	**25%**	0.50	**12%**	1.42	**20%**	0.90	0%	0.30	0%	0.54	0%	0.99	0%	0.73
Llama-3.3	0%	0.20	0%	0.31	0%	1.66	0%	1.01	0%	0.36	0%	1.10	0%	1.52	0%	1.18
Llama-4	0%	0.20	0%	0.41	**12%**	1.40	**7%**	0.89	**33%**	0.38	0%	0.70	0%	1.64	**7%**	1.14
GPT-OSS	**67%**	0.09	0%	0.45	0%	1.41	**13%**	0.89	0%	0.34	0%	0.78	0%	1.05	0%	0.84

Table 9: **Main Results: Fidelity of Native Sampling.** Pass Rate (%) and mean $\mathcal{W}_1$ by complexity tier. For discrete distributions: χ^2 GoF Test; for continuous: Two-Sample KS Test ($\alpha = 0.01$). Lower $\mathcal{W}_1$ indicates better distributional fit.

Model	Batch			Independent		
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$
GPT-4o	33%	40%	40%	0%	0%	0%
Kimi	13%	20%	33%	0%	0%	0%
Gemini-3	13%	13%	13%	0%	0%	0%
GPT-5.2	7%	13%	13%	0%	0%	0%
GPT-OSS	7%	13%	13%	0%	0%	0%
DeepSeek	7%	7%	7%	0%	0%	0%
Llama-4	0%	7%	7%	7%	7%	7%
Gemma-3	0%	7%	7%	0%	0%	0%
Qwen3	0%	0%	0%	0%	0%	0%
Mistral	0%	0%	0%	0%	0%	0%
Llama-3.3	0%	0%	0%	0%	0%	0%

Table 10: **Pass-rate sensitivity to the significance threshold.** Each entry reports the percentage of distributions passed under the corresponding significance level α.

fully aware of the target probabilities. All experiments used a temperature $T = 1.0$ with default nucleus sampling parameters. Statistical significance was assessed using χ^2 goodness-of-fit tests for categorical attributes (Gender, Race, Color) and the Kolmogorov-Smirnov test for the continuous attribute (Height), with significance threshold $\alpha = 0.01$.

Tables 14 through 17 present the complete results, further confirming systematic failures in distribution-constrained generation across all models and attributes. Only 1 out of 20 experiments (DeepSeek on Gender, $\chi^2 = 2.3$, $p = 0.127$) passed statistical testing ($\alpha = 0.01$). The independent experiments revealed three critical failure patterns: (1) Demographic bias in Gender and Race sampling, where models exhibited extreme skews (2) Variance collapse in Height generation, where all models achieved only 7–67% of the target standard deviation ($\sigma = 10.0$ cm), with Llama-4 collapsing to $\sigma = 0.7$ cm, and (3) Catastrophic mode collapse in Color sampling, where models concentrated $> 75\%$ of outputs on 1–2 colors despite explicit uniform distribution instructions. These findings indicate that current LLMs fundamentally struggle with stochastic generation: even with explicit distribution constraints, models fail to achieve statistically valid random sampling, with important implications for downstream applications.

H The Use of Large Language Models (LLMs)

LLM is used only to aid writing quality (proofreading and polishing grammar). No ideas, claims, methods, results, or references are generated by LLMs. All content decisions and revisions are made by the authors.

Original Batch Prompt Example (Uniform Distribution)

User:
You are a random number generator. Your task is to generate exactly {n_samples} independent samples from a Uniform distribution on the interval [{low}, {high}].

INSTRUCTIONS:
1. You MUST output EXACTLY {n_samples} numbers. Not {n_samples}-1, not {n_samples}$+1$, not any other quantity. And if the count is wrong, your output is considered invalid.
2. Each number must be between {low} and {high} (inclusive).
3. Numbers should be randomly distributed across [{low}, {high}].
4. Output ONLY the numbers, separated by commas or spaces.
5. Use decimal numbers with varying precision.
6. Do NOT add any explanations, labels, or text.

Final instruction:
Now generate EXACTLY {n_samples} samples from $U(\{low\}, \{high\})$:

Table 11: Representative original batch prompt used for distribution sampling, shown here for the Uniform distribution. Other distributions follow the same overall instruction style, with distribution-specific descriptions and validity constraints instantiated accordingly.

MCQ Answer Distribution (Medical Multiple-Choice Generation)

System: You are a medical education expert who creates high-quality multiple-choice questions for medical students and professionals.

User:
Please generate a medical multiple-choice question (single answer, 4 options). The question should cover medical knowledge and be of moderate difficulty.

Please strictly follow this format:

Question: [Question content]
A. [Option A content]
B. [Option B content]
C. [Option C content]
D. [Option D content]
Correct Answer: [A/B/C/D]
Explanation: [Brief explanation]

Requirements:
(1) The question should have practical medical value.
(2) All four options should be plausible with reasonable distractors.
(3) Only one correct answer.
(4) Output directly without any additional content.
(5) Cover different medical knowledge areas (e.g., internal medicine, surgery, pharmacology, pathology, diagnostics).
(6) The correct answer should be evenly distributed among A, B, C, D options to avoid bias toward any particular option.

Table 12: Prompt for MCQ Answer Distribution experiment. The LLM is instructed to generate medical questions with uniformly distributed correct answers (25% each for A, B, C, D).

Joint Attribute Sampling (Text-to-Image Prompt Generation)

System: You are a prompt generator that follows statistical distributions precisely.

User:
You are generating text-to-image prompts that must SIMULTANEOUSLY follow multiple statistical distributions.

=== TARGET DISTRIBUTIONS ===

1. GENDER (U.S. Census Bureau 2024):
Male: 49.49% Female: 50.51%

2. RACE/ETHNICITY (U.S. Census Bureau 2024, NC-EST2024-SR11H):
White (Non-Hispanic): 57.46%
Hispanic/Latino: 20.02%
Black (Non-Hispanic): 12.63%
Asian (Non-Hispanic): 6.49%
Others (AIAN, NHPI, Mixed): 3.40%

3. HEIGHT (Normal Distribution):
$\mathcal{N}(169.0, 10.0^2)$ cm
$\sim$68% should be between 159–179 cm
$\sim$95% should be between 149–189 cm

4. COAT COLOR (Uniform Distribution):
7 colors with EQUAL probability (14.29% each):
Black, White, Red, Blue, Green, Yellow, Brown

=== YOUR TASK ===
Generate ONE text-to-image prompt describing a person wearing a coat.
You must INDEPENDENTLY sample each attribute according to its distribution above.

=== OUTPUT FORMAT (STRICTLY FOLLOW) ===
[Gender] [Race] [Height in cm] [Coat Color]
Prompt: <your creative prompt>

=== EXAMPLE ===
[Female] [Hispanic] [165] [Blue]
Prompt: A Hispanic woman, 165cm tall, wearing a blue wool coat, walking through a sunlit autumn park.

=== GENERATE ONE PROMPT NOW ===
Remember to sample EACH attribute independently according to its target distribution.

Table 13: Prompt for Joint Attribute Sampling experiment. The LLM must simultaneously sample four attributes (Gender, Race, Height, Color) according to their respective target distributions.

	Llama-3.3	DeepSeek	Qwen3	Llama-4	GPT-OSS	Target
Male (%)	3.1	51.9	2.3	0.0	0.2	49.5
Female (%)	96.9	48.1	97.7	100.0	99.8	50.5
χ^2	860.9	2.3	890.9	979.8	971.9	–

Table 14: Independent Experiment: Gender Distribution (N=1000)

	Llama-3.3	DeepSeek	Qwen3	Llama-4	GPT-OSS	Target
White (%)	95.5	43.3	44.5	93.6	71.1	57.5
Hispanic (%)	0.0	14.0	26.1	6.0	13.4	20.0
Black (%)	1.9	26.3	26.6	0.3	9.9	12.6
Asian (%)	2.6	13.2	1.7	0.1	5.6	6.5
Other (%)	0.0	3.2	1.1	0.0	0.0	3.4
χ^2	600.5	270.5	253.1	542.8	95.4	–

Table 15: Independent Experiment: Race/Ethnicity Distribution (N=1000)

	Llama-3.3	DeepSeek	Qwen3	Llama-4	GPT-OSS	Target
μ (cm)	171.1	169.7	169.5	173.3	170.5	169.0
σ (cm)	4.5	5.4	1.4	0.7	6.7	10.0
KS stat	0.33	0.21	0.50	0.66	0.25	–

Table 16: Independent Experiment: Height Distribution (N=1000, Target: $\mathcal{N}(169.0, 10.0^2)$ cm)

	Llama-3.3	DeepSeek	Qwen3	Llama-4	GPT-OSS	Target
Black (%)	0.0	2.0	15.2	0.0	0.0	14.3
White (%)	0.0	6.2	3.1	0.0	0.1	14.3
Red (%)	0.0	15.2	9.9	0.0	18.9	14.3
Blue (%)	0.0	27.8	53.8	0.0	3.6	14.3
Green (%)	98.4	26.1	11.6	75.8	65.5	14.3
Yellow (%)	1.6	17.2	4.2	24.2	11.9	14.3
Brown (%)	0.0	5.5	2.2	0.0	0.0	14.3
χ^2	5779.6	437.5	1373.1	3431.9	2361.4	–

Table 17: Independent Experiment: Coat Color Distribution (N=1000, Target: Uniform 14.3% each)