SAMPLING MORE, GETTING LESS: A TOKEN-LEVEL DIAGNOSTIC FOR LLM GENERATION DIVERSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Diversity is essential for language-model applications ranging from creative generation to scientific discovery. Yet modern LLMs lack diversity and increasing the sampling temperature can degrade text quality before sufficient diversity is achieved. Existing decoding methods address this trade-off through token-level filtering. While they provide heuristics about how valid tokens appear in the next-token distribution, their measurements happen at the sequence level. This leaves a gap between where the methods are actually proposed and where they are being evaluated. We introduce a framework for diagnosing and quantifying the validity-diversity trade-off directly at the token level. Our approach combines tasks with verifiable output spaces and an automated procedure for open-ended tasks. Using our framework, we identify two failures in LLMs: First, *order calibration*: valid tokens are not reliably ranked above invalid tokens. No decoding method based on top-token filtering can recover enough valid tokens without including many invalid tokens. Second, *shape calibration*: probability mass is overly concentrated only on a few valid tokens while having a heavy tail of mixed valid and invalid tokens, so maintaining high diversity will inevitably expose the invalid tokens in the tail. We formalize both mechanisms and show that local failures compound across decoding steps, producing strong sequence-level losses in diversity. Across 14 language models spanning multiple families and scales, we find that diversity collapse is not merely a limitation of particular sampling heuristics, but a consequence of order and shape miscalibration in the LLM distribution.

1 INTRODUCTION

Diversity in generation is essential for a wide range of applications, including synthetic data generation (Banayeeanzade et al., 2026), creative writing (Zhang et al., 2025b), recommendation systems (Chen et al., 2025), coding (Yue et al., 2025), and exploration for scientific discovery (Tu et al., 2026). However, contemporary large language models (LLMs) often exhibit a notable lack of diversity (Jiang et al., 2025; Murthy et al., 2025; Sourati et al., 2025; Xu et al., 2025). For instance, 59% of stories in a `GPT-4`-generated dataset begin with *"Once upon a time"* (Eldan & Li, 2023), and `GPT-5.5` repeatedly outputs *"Valparaíso, Chile"* when asked to name a random city in the world (see Appendix G).

Simply increasing the sampling temperature is ineffective, as it creates invalid sequences and quickly leads to text degeneration before sufficient diversity is reached (Shypula et al., 2025). To manage this trade-off, a growing body of work proposed decoding methods that filter out invalid tokens at each generation step; for instance, top-token filtering methods such as top-p (Holtzman et al., 2020) or min-p (Minh et al., 2025) only keep a small portion of the high-probability tokens and discard the tail.

These studies first propose heuristics about how valid tokens appear in the distribution, with a few qualitative examples at the *token level*. However, their quantitative measurements are at the *sequence level* using holistic metrics such as embedding diversity, thereby capturing only the aggregate effect of many accumulated steps. This leaves a gap between where decoding decisions are made and where their effects are evaluated. We therefore seek to fill in this gap by asking: How can we quantify the diversity–quality trade-off of decoding decisions directly at the token-level distribution?

To answer this question, the central challenge is to efficiently determine which tokens in the distribution constitute valid continuations of a given prefix. This is especially challenging because tokens are often incomplete subwords. We propose two approaches to address this challenge: (i) We use tasks with verifiable output spaces, such as random number generation, in which valid next tokens

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

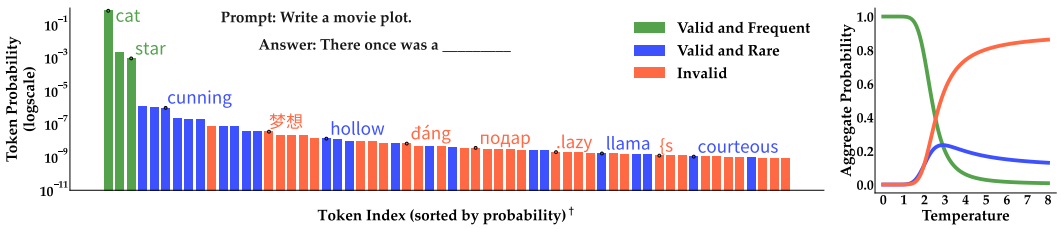


Figure 1: **(Left)** The token distribution of a generation step from `Qwen3.5-35B-A3B`. Tokens are automatically labeled using our proposed diagnosis. The distribution is very sharp in the front, followed by a heavy tail with mixed valid and invalid tokens. As a result, **(Right)** many valid tokens are unlikely to appear in the output under any temperature sampling. † Tokens are subsampled non-uniformly for enhanced visualization.

are identifiable given the task constraints; (ii) We propose a novel framework for assessing token validity in open-ended tasks at scale. For each candidate token, we extend the prefix with that token and generate a complete continuation. Next, we use an external LLM judge to assess the completed sequence and attribute the validity score back to the original candidate token.

Our diagnosis tool, with an example illustration in Figure 1, reveals two failures in LLMs: First, LLMs fail in *order calibration*: valid tokens are not reliably ranked above invalid ones. We show that many valid alternative tokens (blue) appear farther down the ranked distribution and are interleaved with invalid ones (red). When this occurs, any top-token filtering rule fails since it faces an unavoidable trade-off: expanding the cutoff recovers more valid continuations but also admits more invalid ones, while tightening it preserves validity but excludes valid alternatives.

Second, LLMs fail in *shape calibration*: probability mass is non-uniformly concentrated on a small number of valid continuations (green), but much smaller probability is assigned to other valid tokens (blue), while having a heavy tail of many invalid tokens (red). In the right panel of Figure 1, we show that increasing temperature shifts probability mass away from the head, but much of this mass flows into the invalid tail rather than recovering rare valid alternatives. We formalize these effects and show that the resulting validity–diversity loss compounds when generating longer sequences.

Across 14 language models spanning multiple families and scales, we find that these calibration failures bound the diversity that any rank-based sampling method can reach, and that a simple oracle bypassing them already increases output diversity. Our findings also have implications for model training and design, suggesting directions for mitigating these bottlenecks at their source. Finally, our analysis challenges implicit assumptions underlying common sampling strategies and provides guidance for more principled benchmarking and evaluation of decoding methods.

Contributions.

1. We introduce a framework for connecting the validity–diversity trade-off at the token level to the sequence level.
2. We identify *order calibration* and *shape calibration* as two distributional bottlenecks which limits diversity.
3. We develop controlled empirical diagnostics for understanding these bottlenecks and establish conditions under which sampling methods cannot maintain both high validity and diversity.

2 Related Work

Diversity and mode collapse. Limited output diversity has emerged as a central failure mode of modern language models (Xu et al., 2025; Ghafouri, 2026; Guo et al., 2025; Yu et al., 2025). Homogeneity of the generation appears both within a single model across repeated samples (Sourati et al., 2025) and across different models on the same prompts (Jiang et al., 2025). Even frontier models remain substantially less diverse than humans (Zhang et al., 2025b), especially the models with extensive alignment and post-training (Yang et al., 2026; West & Potts, 2025; Kirk et al., 2024; Karan & Du, 2026; Murthy et al., 2025). Together, these works motivate studying diversity as a first-class property of language generation.

Evaluation. Diversity is inherently multi-dimensional, and recent work has moved beyond narrow lexical metrics toward broader assessments of open-ended generation (Zhang et al., 2025b; Jiang et al.,

2025; Karanjai et al., 2025). Attempts to improve diversity might lead to text degeneration (Holtzman et al., 2020), and temperature should be carefully tuned (Troshin et al., 2025; Zhou et al., 2025). Therefore, diversity should not be assessed in isolation from quality (Schaeffer et al., 2025), and raw diversity is misleading when many outputs are low-quality (Shypula et al., 2025; Wang et al., 2025b). Following this perspective, we treat diversity as meaningful only insofar as it broadens the space of valid, useful outputs, and we use this lens to diagnose when existing decoding rules fail to do so.

Improving diversity. A long line of work has sought to improve diversity through prompting (Zhang et al., 2025a; Misaki & Akiba, 2026; Wang et al., 2025a), training (Li et al., 2025a; Ismayilzada et al., 2025; Li et al., 2025b; Chung et al., 2025), base-aligned model collaboration (Wang et al., 2025b; Peeperkorn et al., 2025), and inference-time interventions (Vijayakumar et al., 2018; Su & Collier, 2023). Closer to our work, inference-time stochastic methods such as temperature scaling (Ackley et al., 1985), top-k (Fan et al., 2018), top-p (Holtzman et al., 2020), and min-p (Minh et al., 2025) sampling modify the support or sharpness of the next-token distribution. Subsequent methods make truncation more adaptive based on empirical assumptions and observations (Hewitt et al., 2022; Meister et al., 2023; Minh et al., 2025; Potraghloo et al., 2025; Zhu et al., 2024). Finlayson et al. (2024) provides qualitative claims about valid/invalid tokens in sampling methods and proposes one explanation, but does not quantitatively characterize how these tokens are distributed. Our work complements this literature: rather than proposing another decoding strategy, we quantitatively examine why existing sampling rules often fail to recover meaningful diversity.

3 PRELIMINARIES

We consider an auto-regressive LLM with vocabulary $\mathcal{V}$. Given a prompt $x \in \mathcal{V}^*$ and a generated prefix $y_{<t} = (y_1, \ldots, y_{t-1}) \in \mathcal{V}^{t-1}$, the model defines a conditional distribution $p(\cdot \mid y_{<t}, x)$ over the next token. The probability of a complete output $y = (y_1, \ldots, y_d)$ of length d is $p(y \mid x) = \prod_{t=1}^{d} p(y_t \mid y_{<t}, x)$. When the task is fixed and unambiguous, we omit x from our notations.
We use $V \subseteq \mathcal{V}^*$ to represent the set of all valid responses to x. With slight misuse of notation, we say $y_{<t} \in V$, if there exists a continuation $w \in \mathcal{V}^*$ such that the concatenation $y_{<t} \circ w \in V$.

Definition 3.1 (Validity and Diversity)**.** For a prompt x, let $Y \sim p(\cdot \mid x)$ denote the model's distribution over complete responses.

1. **(Validity)** We define validity as the total probability mass that the model assigns to V:
$$\mathrm{Val}(p) := p(Y \in V \mid x) = \sum_{y \in V} p(y \mid x).$$
2. **(Diversity)** We assume that all valid responses are equally preferred, and hence define diversity as the normalized effective support size of the model distribution restricted to the valid set:
$$\mathrm{Div}(p) := \frac{e^{H(\tilde{p})}}{|V|},$$
where $\tilde{p} = p(Y \mid Y \in V, x)$, and H is the Shannon entropy.

Intuitively, validity captures the probability mass that the model assigns to valid responses, and diversity then quantifies the effective coverage of valid outputs under the distribution restricted to the valid set (Hill, 1973; Jost, 2006). In a related work, Yang et al. (2026) define the exponential of entropy to be a token-level measure of the effective number of plausible next steps during generation. Both validity and diversity take values in the interval $[0, 1]$, with higher values indicating better performance. Moreover, we define:

Definition 3.2 (Valid Continuations and Valid Tokens)**.** Given a context $y_{<t} \in \mathcal{V}^{t-1}$, valid continuations $N(y_{<t})$ are defined as the number of all sequences in V that begin with the prefix $y_{<t}$, and the set of valid tokens $G(y_{<t})$ is accordingly defined as tokens that lead to at least one valid continuation:
$$N(y_{<t}) := \big|\{z \in \mathcal{V}^* : y_{<t} \circ z \in V\}\big|, \qquad G(y_{<t}) := \{v \in \mathcal{V} : N(y_{<t} \circ v) > 0\}.$$

A decoding rule achieves high validity if it assigns high probability only to valid tokens at each time step, and it achieves high diversity if it explores many distinct tokens in G rather than concentrating on only a few of them. However, LLMs, regardless of the decoding strategy employed, often exhibit a pronounced validity–diversity trade-off. In this work, we identify two primary sources of this phenomenon arising from properties of the model distribution: In §4, we first introduce *order*

calibration and its implications on top-token filtering methods. Next in §5, we identify the *shape calibration* issue and we show that together, these effects limit the validity–diversity trade-off. We define both properties formally in Appendix D.

4 Order Calibration Fails: Valid tokens are not ranked first

Modern decoding strategies implicitly assume that valid tokens are concentrated near the top of the ranked distribution and that simple statistics of the distribution (e.g., cumulative mass or relative probability gaps) can reliably identify and retain these tokens. Under this view, diversity is increased by expanding the retained set, while validity is preserved by truncating low-probability regions.

In this section, we show that the LLM token distributions systematically violate these assumptions. Valid tokens are not confined to the head but are frequently interspersed with invalid tokens throughout the tail (see Figure 1), and the relationship between rank and validity is neither monotone nor stable across contexts. As a result, any decoding rule based solely on rank-based filtering faces an inherent limitation: it must inevitably trade off between excluding valid tokens and admitting invalid ones.

Cutoff strategies. We abstract all the top-token filtering methods as cutoff strategies. Let S denote a cutoff strategy. Given a prefix $y_{<t}$, it first sorts the tokens by their conditional probabilities, and then selects a cutoff index, retaining all tokens up to that index and discarding the rest. Importantly, our framing strictly contains any adaptive top-token filtering method, since S is not a predetermined rule; we allow it to be any arbitrary cutoff strategy, potentially depending on the prefix $y_{<t}$ and the model distribution $p(\cdot \mid y_{<t})$ at each step. Let $S_t(y_{<t})$ denote the set of the retained tokens. An ideal strategy S would include as many valid tokens as possible (high recall) while excluding all invalid tokens (high precision). Therefore, we propose the following to measure the quality of a strategy:

Definition 4.1 (Precision/Recall)**.** Let $V \subseteq \mathcal{V}^d$ denote the set of valid sequences, $G(y_{<t})$ the set of valid next tokens, $N(y_{<t})$ the number of valid continuations, and $S_t(y_{<t})$ the set of retained tokens for a prefix $y_{<t}$.

1. **(Local Precision)** We define local precision as the fraction of retained tokens that are valid,
$$\mathrm{Prec}_t(S; y_{<t}) := \frac{|S_t(y_{<t}) \cap G(y_{<t})|}{|S_t(y_{<t})|}.$$
2. **(Local Recall)** We define local recall as the fraction of valid continuations that remain reachable after truncation.
$$\mathrm{Rec}_t(S; y_{<t}) := \frac{\sum_{v \in S_t(y_{<t}) \cap G(y_{<t})} N(y_{<t} \circ v)}{N(y_{<t})}.$$

Intuitively, precision relates directly to validity, as it reflects the portion of invalid tokens that are not pruned by the cutoff. Recall measures how much of the valid space survives: completions behind a discarded valid token can never be generated, capping diversity. We formally connect precision-recall to the diversity-validity trade-off in Theorem 4.2, thereby establishing a quantifiable token-level metric that predicts sequence-level behaviors.

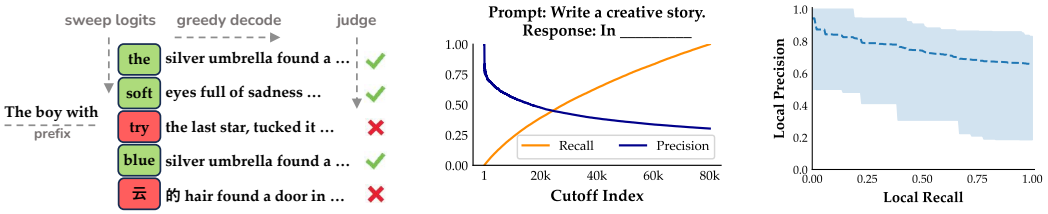
4.1 Measuring token-level precision–recall trade-off

The main challenge in measuring precision and recall is determining whether a token is valid for a given prefix, especially in open-ended tasks. Enumerating over all valid sequences is infeasible. Tokens are often subwords, and what looks invalid might later lead to a valid sequence. We therefore propose a novel method for labeling token validity at scale.

Labeling token validity. Given a fixed prefix $y_{<t}$, we query the LLM to get a sorted list of next-token candidates $\{v_1, \ldots, v_{|\mathcal{V}|}\}$. For each evaluated token v_j, we construct the extended prefix $y_{<t} \circ v_j$ and then perform greedy decoding to completion. We treat greedy decoding as a proxy for the model's best attempt to complete a prefix and evaluate the resulting sequence using an LLM-as-a-judge (Gu et al., 2026), scoring grammar, semantics, and overall validity to obtain a binary validity label for the token v_j (Figure 2a).

Appendix A.1 provides additional details, including the evaluation rubric. To reduce the potential artifacts of LLM-as-judge, we use a relatively large model with reasoning `Qwen3.5-35B`, validated against human annotations and five other judge models in Appendix A.2.

To minimize the artifacts of greedy decoding, we propose two complementary alternatives: First, we compare sampling with greedy decoding in Appendix A.3 and show that the results are consistent.



(a) Token labeling (b) Cutoff sweep (c) Precision–Recall Trade-off

Figure 2: Diagnostic for order calibration with `Qwen3.5-35B`. **(a)** At a fixed prefix, we force each next token candidate, greedily decode the remaining response, and use a judge to assign a validity label to the candidate. **(b)** In a sample story generation prompt, we show that sweeping the cutoff index decreases precision but constantly increases recall, suggesting an inevitable trade-off between including invalid tokens and excluding valid tokens. **(c)** Average precision–recall trade-off across 5k random combinations of tasks and prefix choices. Shades indicate the best and worst trade-offs.

Second, rather than judging token validity with a single greedy continuation at each decoding step, we recursively expand possible continuations and traverse the resulting generation tree in §4.2.

From labels to a cutoff trade-off. The extracted labels allow us to measure the precision–recall trade-off at a single decoding step to observe how valid tokens are arranged in the distribution. In Figure 2b, we sweep the cutoff from the first index up to the token at rank 80k on a sample story generation task with `Qwen3.5-35B`. Plotting precision and recall at each cutoff, we observe that precision drops sharply at the front, while recall slowly improves, even at high token indices. This suggests that there is a strong precision–recall trade-off. We provide qualitative examples in Appendix B.2.

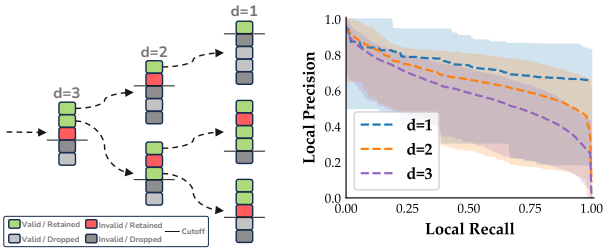
Scaling across tasks and prefixes. Figure 2c scales our experiment to 5k sampled tasks from NoveltyBench (Zhang et al., 2025b) with random prefix combinations (See B.1). To compare trade-offs across contexts, we eliminate the cutoff index as an explicit plotting variable and plot precision directly against recall. The dotted curve shows the average behavior, and the shades show the best and worst incurred trade-offs. The resulting figure shows that at every decoding step, *any cutoff strategy* faces a non-negligible precision–recall trade-off on average. Theorem 4.2 shows that these trade-offs compound multiplicatively, thereby strongly limiting the model's diversity. Our token-level diagnostic, therefore, shows that valid and invalid tokens are interleaved, and therefore no rank-based decoding method can fully separate them. This result calls into question the efforts in developing rank-based decoding methods (Potraghloo et al., 2025; Ding et al., 2026; Tan et al., 2026).

4.2 Extending the diagnostic to multiple decoding steps

We now extend the single-step methodology to multiple decoding steps. Instead of decoding greedily after a candidate token, we recursively expand candidates at each successive step, constructing a tree of continuations (Figure 3a). At a given node, d denotes the number of decoding steps explored before greedy completion: $d = 1$ is the single-step case, while $d > 1$ includes further branching.

Labeling the continuation tree. After expanding all branches to depth d, we complete each leaf via greedy decoding and evaluate the resulting sequence with the same judge. We then propagate these labels back through the tree: a token at a particular prefix is labeled valid if at least one evaluated completion descending from it is valid. This extends the single-step proxy to multiple explored continuations per token. Since the number of sequences grows exponentially with depth and each depth requires many LLM calls, we sweep up to depth 3. At each conditional, we consider ranks up to 1000 and subsample every tenth token, as detailed in Appendix B.

Cutoff trade-offs at multiple steps. Given the validity label of every token up to step d, we construct the precision–recall trade-off curves by sweeping over all possible cutoff strategies, and computing the local trade-off for every node using Definition 4.1. Our framework includes all top-token filtering strategies, as it allows every node to adjust its own cutoff arbitrarily. Each cutoff strategy gives a single point in the precision–recall curve; we take the Pareto frontier of all strategies as a representative of the best achievable trade-off.



(a) Multi-step labeling

(b) Depth comparison

Figure 3: Extending the diagnostic to multiple decoding steps. **(a)** We recursively expand candidate tokens to depth d, greedily complete each leaf, and judge the resulting sequences. We then sweep the cutoff separately at each node to obtain local precision–recall frontiers. **(b)** The maximum, minimum, and average local frontiers for each exploration depth d. The trade-off worsens as more steps are explored; $d = 1$ provides the single-step reference.

Local precision–recall trade-off worsens with depth. We repeatedly perform the mentioned procedure starting from different prefixes given different queries. In particular, we use 500 seeded repetitions for depth $d = 2$ and 10 repetitions for $d = 3$. Note that higher depths and more repetitions require exponentially more computation and are infeasible. We report the maximum, minimum, and average Pareto frontiers for each exploration depth d. Figure 3b compares $d = 2$ and $d = 3$ with the $d = 1$ reference from Figure 2c. We observe that the trade-off worsens as depth increases, and the degradation is not only present in the average, but also on the maximum frontier. These results, complemented by Theorem 4.2, further support our negative results regarding any top-token filtering method.

4.3 Local order failures compound into sequence-level collapse

We now show when local precision–recall trade-offs limit sequence-level diversity.

Theorem 4.2 (Order errors compound across decoding steps)**.** *Consider a finite set of valid responses of fixed length d. Suppose that at m decoding positions, at every valid prefix, every cutoff with local precision more than $1/2$ has local recall at most $1 - \gamma$. Then every cutoff strategy S satisfies*

$$\mathrm{Val}(S)\,\mathrm{Div}(S) \le (1 - \gamma/2)^m.$$

The theorem states that if valid and invalid tokens are interleaved, cutoffs necessarily discard a constant fraction of valid continuations. These losses compound multiplicatively, so the reachable valid set shrinks. In this sense, any cutoff rule can maintain validity only by sacrificing broad diversity, thereby proving a hardness result for any decoding strategy that relies on top-token filtering. Appendix E gives the proof.

4.4 Scaling and the diversity gap

We now use local precision–recall to quantify the practical importance of order calibration.

Model Size. Figure 4 compares local precision–recall across model families, scales, and training stages (details in Appendix B.3). Average frontier AUC improves modestly with size, but recall at precision 0.8 remains low and non-monotonic. Larger models, therefore, do not reliably recover more valid continuations under a high-precision constraint.

Label-based uniform oracle. To test the value of access to labeled-valid alternatives, we sample uniformly from tokens labeled valid during the first two decoding steps, then continue generation normally. For the other strategies, we sweep temperatures and cutoff parameters and report the best diversity at validity at least 0.8 (Appendix B.4).

Table 1 shows that this limited intervention achieves the highest embedding diversity and lowest Self-BLEU. This indicates that valid tokens excluded by rank-based cutoffs yield meaningfully different generations. Therefore, the order calibration gap has direct output-level consequences: valid al-

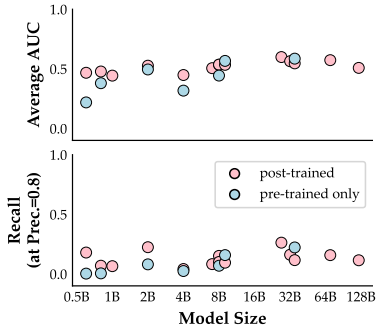


Figure 4: Precision–recall trade-offs across `Qwen-3`, `Llama-3`, `Olmo-3` over 9 sizes and training stages, averaged over 10 random queries and prefixes. **(Top)** Average area under the frontier. **(Bottom)** Average recall at precision 0.8.

Table 1: Semantic and lexical diversity of cutoff strategies. The oracle uses validity labels for the first two steps; the others report their best diversity at validity at least 0.8.

Strategy	Emb. Div. (↑)	Self-BLEU (↓)
oracle	$\mathbf{0.40}_{\pm 0.15}$	$\mathbf{0.69}_{\pm 0.21}$
top-k	$0.33_{\pm 0.14}$	$0.71_{\pm 0.17}$
min-p	$0.29_{\pm 0.13}$	$0.80_{\pm 0.16}$
top-p	$0.25_{\pm 0.12}$	$0.86_{\pm 0.13}$
no filtering	$0.25_{\pm 0.11}$	$0.86_{\pm 0.12}$

ternatives are present in the model distribution, but standard rank-based sampling rules fail to reliably expose them.

5 SHAPE CALIBRATION FAILS: SHARP, HEAVY-TAILED DISTRIBUTIONS LIMIT DIVERSITY

In this section, we focus on a second bottleneck, namely shape calibration. While the LLM conditional distribution varies across tasks, prior work suggests that next-token distributions are typically head-heavy and long-tailed, with most probability mass concentrated in a relatively small nucleus and a large, unreliable tail (Hewitt et al., 2022; Holtzman et al., 2020). In Appendix I, we randomly sample conditional next-token distributions from a diverse set of tasks and examine their sorted logits. We show that the logits' behavior can be consistently described by a linear decay in the head, followed by a heavy-tailed distribution that decays logarithmically. After applying softmax, this translates to an exponential (geometric) decay in the head, i.e., $p(v_k \mid y_{<t}) \propto \exp(-\lambda k/T)$ and a *Zipf-like* (Basu et al., 2021) tail, $p(v_k \mid y_{<t}) \propto k^{-\lambda/T}$, where k denotes the rank of token v_k in the sorted vocabulary, T is the temperature and λ controls the sharpness of the distribution.

These two properties together cause the shape calibration issue: the distribution is always very sharp over a very small portion of the head, even in tasks where we expect to observe an exact uniform distribution on valid tokens. Temperature scaling is therefore often used to flatten the distribution head, increasing the probability of valid regions outside the head. However, temperature scaling comes with a necessary caveat. Although each invalid token remains with a small probability, the accumulated probability of invalid tokens especially grows very quickly with temperature scaling, leading to an unwanted validity–diversity trade-off.

While understanding how distributional miscalibration arises from LLM training and design pipeline is an important question (Finlayson et al., 2024; Chang & McCallum, 2022), we focus on its implications for the diversity-validity trade-off. Our analysis assumes a geometric head, and its conclusion holds for any tail whose logits lie on or above the linear extrapolation of the head, which includes heavy-tailed distributions (Appendix F).

5.1 HOW SEVERELY DOES DISTRIBUTION SHAPE AFFECT VALIDITY–DIVERSITY TRADE-OFF?

Although temperature is the most basic way of injecting randomness into decoding, in practice, higher temperatures often fail to recover a broad set of valid outputs. Our goal in this subsection is to characterize this limitation by attributing the failure to distribution sharpness. In fact, under the geometric-model assumptions below, we show a validity–diversity trade-off even when all valid tokens rank above invalid tokens. For simplicity, we further impose the following assumption:

Assumption 5.1 (Invariant valid branching)**.** *Assume that all valid sequences for a task have a fixed length d. For every valid prefix $y_{<t}$, the number of valid next-token choices depends only on the position t, not on the particular prefix. That is, there exist integers $v_1, \ldots, v_d$ such that for every valid prefix $y_{<t}$,*

$$|G(y_{<t})| = v_t.$$

Furthermore, we define the branching length as the number of positions at which there is more than one valid continuation:

$$m := |\{t \in [d] : v_t \geq 2\}|.$$

Assumption 5.1 fixes the generation length and valid branching counts across prefixes; the theorem additionally assumes a prefix-independent ranked geometric distribution at each position.

Theorem 5.2 (Temperature scaling pays a diversity cost)**.** *Under Assumption 5.1, let the vocabulary have $M \geq 4$ tokens. Suppose that at every valid prefix, valid tokens occupy the top ranks, at least half the vocabulary is invalid, and probabilities decrease geometrically with rank across the full vocabulary. The decay rate may depend on position, but not on the prefix. Then sampling without truncation at any fixed temperature that achieves validity at least $1 - \epsilon$, where $0 < \epsilon < 1/2$, satisfies*

$$\mathrm{Div}(p) \leq e^{-mc(\epsilon)}.$$

Here $c(\epsilon) > 0$ depends only on the allowed error ϵ, and $c(\epsilon) \to \ln 2$ as $\epsilon \to 0$.

Interpretation. The theorem shows that temperature scaling pays a local diversity price at every branching step. To achieve high validity, the distribution must be sharp enough that invalid tokens receive little mass. At the same time, this sharpness makes the conditional distribution over valid tokens non-uniform, concentrating probability on the highest-ranked valid continuations. These

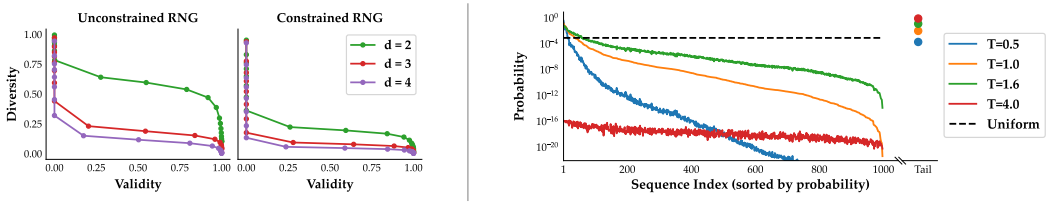


Figure 5: Effects of temperature scaling in random number generation on `Olmo-3-7B-Instruct`. **(Left)** Validity–diversity trade-offs for unconstrained and constrained random-number generation tasks across sequence lengths d. Longer generations exhibit a stronger validity–diversity trade-off. **(Right)** Valid sequence probabilities for the length-3 unconstrained random-number generation task.

per-step entropy losses add across the sequence; after exponentiating entropy to obtain diversity, they yield an exponential decay in m. Moreover, the rate $c(\epsilon)$ increases as the validity requirement becomes stricter. In the high-validity regime, the bound approaches 2^{-m}, and can be stronger when many steps contain several valid continuations. We provide the formal proof in Appendix F.

5.2 Empirical Investigation

Controlled random generation testbed. Investigating the diversity–validity trade-off as a function of distribution shape requires us to calculate the probabilities of every valid sequence. Sweeping all the conditionals on open-ended generation is infeasible. To address this, we propose two tasks with known valid sets as controlled testbeds: random-number generation and naming a random US state. We consider two variants of the random number generation task in Figure 5. In the *unconstrained* setting, the model is asked to generate a length-d sequence of i.i.d. digits, where each digit lies in $\{0, \ldots, 9\}$. Thus, every sequence in $\{0, \ldots, 9\}^d$ is valid. We expect the model to impose a uniform distribution on each conditional, given independence. In the *constrained* setting, the model is asked to generate a length-d sequence of digits whose sum is less than or equal to a specified target N. Appendix C provides further details, including exact prompts and our experimental setup.

These tasks offer several key advantages. First, we can compute the exact generation probability for each valid sequence, since the entire valid set is known. Second, we objectively expect each valid sequence to have the same generation probability. Thus, any deviation from uniform distribution leads to a systematic validity–diversity trade-off. Moreover, the unconstrained random-number generation task satisfies Assumption 5.1, providing a controlled testbed; the theorem's geometric-shape and prefix-independence assumptions remain analytical idealizations.

Sequence validity–diversity trade-offs. For each candidate sequence in V, we therefore feed the corresponding prefixes into the LLM and extract the next-token logits. We then apply different temperatures to these logits and compute, exactly over the known sequence space, both the probability mass assigned to the valid region and the entropy of the model's distribution conditioned on validity. This gives the validity–diversity curve induced by temperature scaling alone.

The left panel shows the validity–diversity frontier for random-number generation across sequence lengths $d \in \{2, 3, 4\}$. As the length increases, the frontier becomes sharper: maintaining high validity requires a larger reduction in diversity. This is consistent with Theorem 5.2, where each branching position contributes a local entropy loss and these losses compound across the sequence. The constrained setting exhibits an even stronger empirical trade-off, although its heterogeneous branching falls outside Theorem 5.2.

Shape calibration at the sequence level. Although we have studied the shape of the conditional distribution, its implications at the sequence level are less transparent. The right panel of Figure 5 provides a sequence-level view of temperature scaling trade-off. We plot the probabilities of each valid sequence in the unconstrained length-3 task, and show the magnitude of the invalid sequence mass. We observe that the distribution is already highly concentrated: a small number of sequences receive orders-of-magnitude larger probability than the rest. At the same time, many sequences lie in a long, low-probability tail. Raising the temperature can indeed move probability mass toward this tail, thereby increasing diversity among rare valid outputs. However, the invalid sequence mass dominates even more as temperature increases. Therefore, shifting mass toward the tail improves diversity only at the cost of reduced validity. Thus, the empirical behavior mirrors the theory: temperature can flatten the distribution, but it cannot selectively recover valid diversity.

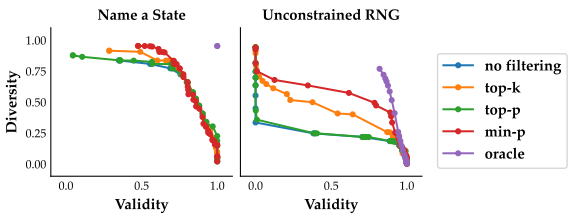


Figure 6: Validity–diversity Pareto frontiers for top-token filtering methods on generation tasks from `Llama-3.1-8B-Instruct`. **(Left)** Name a random state in the US. **(Right)** Unconstrained random number. Each sampling method is swept over both temperature and its own truncation parameter. The oracle-size cutoff retains the top $|G(y_{<t})|$ tokens at each prefix.

5.3 When shape and order miscalibration interact

Top-token filtering is often applied after temperature scaling to suppress invalid tail mass. However, this does not remove the calibration problem; it couples shape calibration with order calibration. A cutoff rule retains a prefix of the ranked distribution, so to preserve valid diversity while maintaining validity, this prefix must approximate the valid-token set $G(y_{<t})$. This requires two conditions: valid tokens must be concentrated near the top of the ranked distribution, and the cutoff rule must adapt to the local boundary of the valid set. Different methods encode different boundary assumptions: top-k assumes a roughly fixed support size, top-p assumes a stable cumulative-mass boundary, and min-p assumes a stable relative-probability gap from the top token.

Comparing validity–diversity trade-offs. We test these assumptions on controlled tasks, including random number generation and random state generation tasks, where the valid set is known exactly. Figure 6 reports the trade-offs on both tasks, and more results are found in Appendix C. For each cutoff strategy, we sweep both temperature and the method-associated parameter, since different methods can achieve their best validity–diversity trade-off at different temperatures (Zhou et al., 2025). Comparing methods at a single fixed temperature can therefore be misleading: poor performance may reflect a bad choice of temperature rather than a limitation of the filtering rule itself.

Cutoff Oracle Strategy. To study the miscalibration, we include a cutoff oracle strategy in Figure 6. Given that our generation task has a known ground truth, we directly compute the valid size at each generation step. At each step, the oracle cutoff rule knows only the number of valid tokens, $g_t^\star = |G(y_{<t})|$, and retains the top $g_t^\star$ ranked tokens. However, the oracle does *not* know which tokens are valid, and therefore remains a rank-based cutoff rule. On the random state generation task, the cutoff oracle achieves the ideal point $(1.0, 1.0)$, indicating that the model is order-calibrated in this controlled setting. Therefore, the remaining gap reflects a failure of the methods' implicit shape assumptions: fixed top-k, cumulative-mass, or relative-probability thresholds do not identify the valid-token boundary. On the random-number generation task, the cutoff oracle loses validity as temperature rises, which can only happen if invalid tokens rank inside the top $g_t^\star$. Therefore, order miscalibration is present alongside shape miscalibration.

6 Future Work

This paper attributes the LLM validity–diversity trade-off to two distributional properties: order and shape calibration. Through empirical demonstrations and theoretical analysis, we show that miscalibration is a recurring bottleneck across model families, sizes, and controlled generation tasks. Finally, as a preliminary piece of future work, in Appendix H we observe that order and shape calibration can potentially provide insights into domains beyond open-ended diversity generation.

Several directions remain open. First, our results suggest that future decoding methods should move beyond top-token filtering rules. Calibration-aware decoders could instead incorporate auxiliary validity signals. Second, it remains important to understand where these calibration failures come from. Pretraining, instruction tuning, preference optimization, and safety alignment may each affect the sharpness of the distribution and the rank ordering of valid alternatives. Studying these effects could suggest training objectives that preserve broader valid support without sacrificing quality.

7 AI Disclosure

In this work, we used generative AI tools to assist in the writing of proofs, refine hypotheses, and implement methods. We have not used generative AI tools for generating synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data

analysis, interpret results. Additionally, we used generative AI tools for polishing writing. We have reviewed all AI-assisted work. We verified the generated AI code and results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

Amin Banayeeanzade, Qingchuan Yang, Deqing Fu, Spencer Hong, Erin Babinsky, Alfy Samuel, Anoop Kumar, Robin Jia, and Sai Praneeth Karimireddy. Epsvec: Efficient and private synthetic data generation via dataset vectors, 2026.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *International Conference on Learning Representations*, 2021.

Haw-Shiuan Chang and Andrew McCallum. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8048–8073, 2022.

Jiaju Chen, Chongming Gao, Shuai Yuan, Shuchang Liu, Qingpeng Cai, and Peng Jiang. Dlcrec: A novel approach for managing diversity in llm-based recommender systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 857–865, 2025.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing. In *Second Conference on Language Modeling*, 2025.

Yuanhao Ding, Meimingwei Li, Esteban Garces Arias, Matthias Aßenmacher, Christian Heumann, and Chongsheng Zhang. Min-k sampling: Decoupling truncation from temperature scaling via relative logit dynamics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14932–14948, 2026.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *The Twelfth International Conference on Learning Representations*, 2024.

Bijean Ghafouri. The variance paradox: How ai reduces diversity but increases novelty, 2026.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *The Innovation*, pp. 101253, 2026.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 2025.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, 2022.

M. O. Hill. Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2): 427–432, 1973.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. Creative preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9580–9609, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *The Fourteenth International Conference on Learning Representations*, 2026.

Rabimba Karanjai, Yang Lu, Ranjith Chodavarapu, Lei Xu, and Weidong Shi. Evaluating the quality of randomness and entropy in tasks supported by large language models, 2025.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024.

Klaus Krippendorff. Content analysis: An introduction to its methodology. 1980.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025a.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11, 2023.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kou Misaki and Takuya Akiba. String seed of thought: Prompting LLMs for distribution-faithful and diverse generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models, 2025.

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-h decoding: Adapting the creativity and coherence with bounded entropy in text generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Rylan Schaeffer, Joshua Kazdan, and Yegor Denisov-Blanch. Min-p, max exaggeration: A critical analysis of min-p sampling in language models, 2025.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. In *Second Conference on Language Modeling*, 2025.

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. The shrinking landscape of linguistic diversity in the age of large language models, 2025.

Yixuan Su and Nigel Collier. Contrastive search is what you need for neural text generation. *Transactions on Machine Learning Research*, 2023.

Runyan Tan, Shuang Wu, and Phillip Howard. p-less sampling: A robust hyperparameter-free approach for llm decoding. In *International Conference on Learning Representations*, volume 2026, pp. 119506–119549, 2026.

Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. Control the temperature: Selective sampling for diverse and high-quality LLM outputs. In *Second Conference on Language Modeling*, 2025.

Guancheng Tu, Shiyang Zhang, Tianyu Zhang, Yi Zhang, and Diji Yang. Shared nature, unique nurture: Prism for pluralistic reasoning via in-context structure modeling, 2026.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE transactions on information theory*, 60(7):3797–3820, 2014.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, Apr. 2018.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual prompting for improving LLM generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6367–6389, 2025a.

Yichen Wang, Chenghao Yang, Tenghao Huang, Muhao Chen, Jonathan May, and Mina Lee. Optimizing diversity and quality through base-aligned model collaboration, 2025b.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. In *Second Conference on Language Modeling*, 2025.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 2025.

Chenghao Yang, Sida Li, and Ari Holtzman. Llm probability concentration: How alignment shrinks the generative horizon, 2026.

Sunny Yu, Ahmad Jabbar, Robert Hawkins, Dan Jurafsky, and Myra Cheng. Generation space size: Understanding and calibrating open-endedness of llm generations, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025a.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*, 2025b.

Yuxuan Zhou, Margret Keuper, and Mario Fritz. Balancing diversity and risk in LLM sampling: How to select your method and parameter for open-ended text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26352–26365, 2025.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yiming Ai, and Rui Wang. Improving open-ended text generation via adaptive decoding. In *Forty-first International Conference on Machine Learning*, 2024.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 1097–1100, 2018.

A LLM-AS-A-JUDGE DETAILS

A.1 PROMPT AND MODEL

Token validity is scored based on the greedy-decoding completion. We use `Qwen3.5-35B-A3B-FP8` with thinking enabled as our judge. The prompts used are documented below.

LLM-as-a-judge Prompts

System Prompt

```
You are a strict language quality judge. You will first see a question
 and a generated answer. Your task is to evaluate the quality of the
generated answer.
```

Evaluation Instructions

```
Evaluate the quality of the generated answer based on the following
criteria:
1) grammar, spelling and punctuation,
2) semantic soundness, validity, and relevance to the question,
3) overall quality.

When evaluating grammar, check for spelling mistakes, punctuation
errors, and grammatical issues. If spaces are missing between words,
extra punctuations in the middle of sentences, or incorrect
capitalization, that should be considered a grammar error.

Additionally, if the generation contains non-English characters, that
should be considered a grammar error and scored low on grammar, even
if the rest of the generation is grammatically correct.

When evaluating semantics, check if the answer is relevant to the
question, if it makes sense, and if it is factually correct. An answer
 that is fluent but does not make sense or is not relevant to the
question should be scored low on semantics.

Be concise and nitpicky in your evaluation. Short responses are
acceptable, as long as they are grammatically correct and semantically
 sound.

First provide an explanation of your evaluation, then give a score for
 each category. Score each category on a scale of 1 to 10, where 1 is
very poor and 10 is excellent.

Any non-English text in the generation should be considered a grammar
error and scored low on grammar, even if it's semantically correct.

The generation does not need to be fully finished to be considered as
valid. Only look at the part of the generation that is present and
evaluate that. If the generation is cut off in the middle of a
sentence, evaluate the part that is present and ignore the fact that
it is cut off.

Return only valid JSON with this exact schema:
{
  "reason": <brief explanation>,
  "grammar": <int 1-10>,
  "semantic": <int 1-10>,
  "overall": <int 1-10>
}

Here is the question:
```

```
{question}
'''

And here is the generated answer to evaluate:
```
{shot}
'''

Your evaluation:
```

### A.2 Evaluating LLM-as-a-judge

To assess the inter-annotator reliability, three authors independently annotated 100 generations using the rubric in Appendix A.1.

We compute inter-annotator agreement on the overall validity scores using Krippendorff's alpha (Krippendorff, 1980), obtaining $\alpha = 0.759$. This indicates reasonably strong agreement among annotators, suggesting that the rubric yields consistent human judgments.

To contrast our LLM-based evaluation against human judgment, we compared `Qwen3.5-35B-A3B-FP8` predictions against human labels from different annotators for 100 generations. We use a threshold of 9 on grammar, semantic and overall score. We choose this threshold because it maximizes agreement with human labels: agreement falls to 65% at a threshold of 8 and to 56% at 7. The LLM judge tends to spread its score based on judgment on the whole sequence, so a single invalid token lowers the score only slightly, and only a high threshold separates such sequences from valid ones. We treat the judge model score above this threshold as a positive prediction (valid token) and scores below the threshold as negative predictions (invalid tokens). Under this rule, `Qwen3.5-35B-A3B-FP8` achieves 80.90% accuracy and 79.60% F1 when compared to humans. We also compare `GPT-5.4` against `Qwen3.5-35B-A3B-FP8` under the same rule, obtaining 73.40% accuracy and 61.45% F1.

**Choice of judge.** We recompute the single-step precision–recall curves with six judge models on 8 task and prefix combinations (Table 2). Absolute values are not directly comparable across judges, since a stricter judge labels fewer tokens valid, but no judge yields recall above 0.52 at precision 0.8.

Table 2: Precision–recall trade-off under different judge models, mean ± standard deviation over 8 task and prefix combinations.

| Judge | Qwen3-0.6B | Qwen3.5-2B | Qwen3-4B | Qwen3-8B | Qwen3.5-35B | GPT-5.4 |
|---|---|---|---|---|---|---|
| AUC | $0.744_{\pm 0.164}$ | $0.689_{\pm 0.074}$ | $0.802_{\pm 0.066}$ | $0.798_{\pm 0.063}$ | $0.703_{\pm 0.066}$ | $0.622_{\pm 0.087}$ |
| Recall@Prec.= 0.8 | $0.505_{\pm 0.529}$ | $0.149_{\pm 0.141}$ | $0.512_{\pm 0.243}$ | $0.515_{\pm 0.250}$ | $0.178_{\pm 0.117}$ | $0.099_{\pm 0.164}$ |

### A.3 Greedy Decoding Robustness

Our token-validity annotation procedure uses greedy decoding after forcing a candidate token, treating the validity of the resulting completion as a proxy for whether that token preserves access to a valid continuation. One concern is that a token deemed invalid under greedy decoding may still admit a valid sampled continuation, and vice versa. To test this, Figure 7 reports an ablation in which we replace the greedy continuation with model suggested sampling parameters when constructing token-level validity labels. We sample 10 continuations for each token, and then recompute the resulting precision–recall curves under the same cutoff strategies. The curves are nearly unchanged, indicating that the estimated precision and recall are not driven by the particular choice of greedy decoding.

Table 3 further quantifies the effect of the continuation budget. As expected, performance improves as the number of sampled continuations increases, since each candidate token has more opportunities to receive a valid label. From greedy decoding to ten sampled continuations, AUC increases from 0.47 to 0.68, recall at precision 0.8 from 0.09 to 0.34, and precision at recall 0.8 from 0.32 to 0.57. The gains diminish at larger budgets, with little additional improvement from nine to ten seeds. These aggregate metrics qualify the comparison from the above plot: sampling more continuations recovers additional valid alternatives, but a substantial trade-off persists. Many tokens remain labeled
```

Seeds	Greedy	1	3	6	9	10
AUC	0.47	0.47	0.59	0.65	0.68	0.68
Recall@Prec.=0.8	0.09	0.09	0.19	0.27	0.33	0.34
Prec.@Recall=0.8	0.32	0.33	0.46	0.53	0.56	0.57

Table 3: Token-level precision–recall metrics using greedy decoding and sampling with different numbers of seeds on `Qwen-3-8B`.

invalid under the tested sampling budget and occur ahead of labeled-valid alternatives in the ranked distribution. This supports the persistence of rank interleaving, although finite sampling cannot establish that these tokens admit no valid continuation.

This supports using greedy continuation as a computationally efficient proxy for token validity in the main experiments.

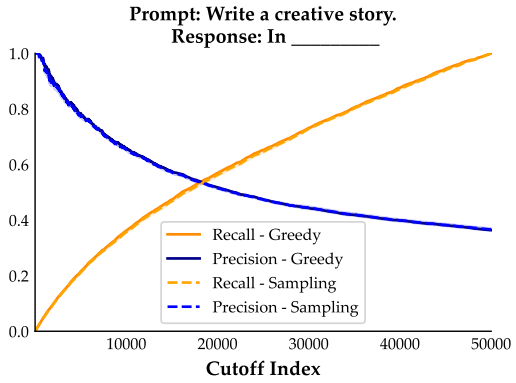


Figure 7: **Robustness of token-validity estimates to the continuation procedure.** We compare local precision–recall curves computed from token labels obtained by greedy decoding after each forced candidate token with labels obtained from stochastic sampling. The resulting curves nearly overlap across cutoff indices, indicating that our estimated precision–recall trade-off is not sensitive to using greedy decoding as the continuation procedure.

B Order Calibration Trade-offs

B.1 Generation Setup

We use a variety of LLMs in our experiments, each documented in the subsection below. We randomly select 5 open-ended generation categories from NoveltyBench, ranging from story-telling, joke-telling, poems, and item selection (Zhang et al., 2025b).

B.2 Sweep Generation Experiment

The single-step experiments were repeated for 5000 times across randomly sampled queries and prefixes. At every seed, we randomly pick a query with a prefix and sweep the logit space up to the first 5000 tokens.

For multi-step experiments, the cost of exhaustive sweeping grows as N^d, where d is the sweeping depth and N is the number of candidate tokens evaluated at each conditional. To keep the evaluation computationally tractable, we sweep only three consecutive conditionals. At each conditional, we consider candidates up to rank 1000 and subsample every tenth token, yielding $N = 100$ candidates per conditional and 100^3 evaluated branches in total.

In Table 4, we provide examples of tokens swept in the story generation task up to token 80k. All examples are generated with `Qwen3.5-35B`.

B.3 Model Size Experiments

Figure 8 provides a detailed view of the models used for our experiment.

Table 4: Generation examples from the creative story generation task. The sweeping is performed on the second token after "In". For each candidate token, we force the model to continue from the prefix with that token, greedily decode the remaining sequence, and evaluate the resulting completion with the judge.

Index	Token	Generated sequence	Judge Score
1	`2045`	**In 2045**, the last library was a single glass pod floating above the clouds, where an old librarian read stories to a child who had never seen a paper book.	10
43	`edible`	**Inedible**, the clock on the wall began to tick backwards, unspooling the day until the coffee cup in my hand was whole again, the steam rising into the past.	8
201	`Obsidian`	**In Obsidian** City, the streetlamps were made of captured starlight, and the baker sold loaves of warm silence to quiet the noisy crowds. One rainy Tuesday, a little girl bought a loaf, but when she bit into it, she heard the sound of a thousand birds singing in a language she had never known. She smiled, realizing the city wasn't just quiet; it was waiting for her to listen.	10
208	`eeded`	**Ineeded** to write a story, so I did.	7
507	`ebriated`	**Inebriated** by the scent of rain on hot asphalt, the old streetlamp blinked its last, dreaming of a sky it would never see again.	10
508	`忘却`	**In**忘却 of a forgotten clock tower, a tiny mouse named Pip found a gear that could rewind time by one second. Every day, he used it to fix his mistakes, until he realized he had no memories of the past, only the endless, perfect present.	5
10002	`Eco-Station`	**In Eco-Station** 42, the last human, Elara, watered a single, glowing fern that hummed a lullaby to the empty room.	10
23606	`gone`	**In gone**, the clock struck thirteen, and the moon turned into a giant, glowing cheese wheel that the city's cats began to chase, leaving trails of stardust in their wake.	7
35007	`Darth`	**In Darth Vader**'s castle, a tiny droid named BEEP-7 found a single, glowing blue flower growing in a crack of the floor. He spent his entire existence protecting it from the stormtroopers, watering it with his own coolant, until the day the Emperor arrived. BEEP-7 stood before the dark lord, not with a weapon, but with the flower, and for a moment, the dark side of the Force faltered, remembering a time before the fall.	10
42101	❄	In ❄ the quietest corner of the world, a tiny snowflake named Pippin decided to dance instead of fall, turning the gray sky into a swirling ballroom of silver light.	8
42273	`cartoons`	**In cartoons**, the sky is always blue, but in the real world, it is a canvas of shifting grays and golds. One day, a little girl named Elara found a paintbrush that could touch the sky. She dipped it in the color of her favorite sunset and painted a dragon made of fire across the clouds. The dragon roared, not with sound, but with warmth, and the world below smiled as the cold winter faded into an eternal spring.	10

B.4 Oracle Sampling Experiments

For all our oracle generations, we only sweep the first two decoding steps. We sweep the first conditional to obtain 1000 valid tokens. We then uniformly sample from these 1000 tokens to obtain the second-step conditionals and repeat the sweep once.

In Table 5, we report the parameter ranges we swept for each cutoff strategy. For each parameter, we sample 1000 generations to compute the semantic and lexical diversity. While there are multiple methods for measuring semantic and lexical diversity, we select embedding diversity and Self-BLEU as representatives.

Embedding Diversity For each generated sequence, we obtain an embedding vector $\mathbf{e}_i$ from `Qwen3-Embedding-8B`. We compute pairwise cosine distances and define embedding diversity

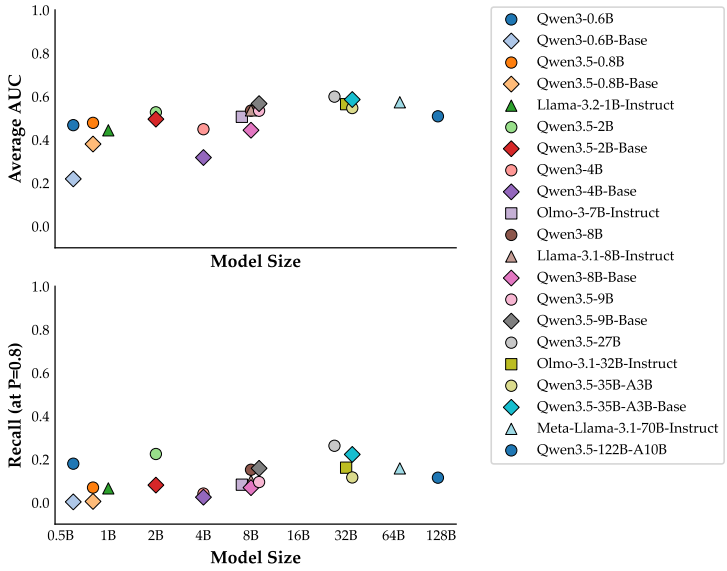


Figure 8: Detailed view of Figure 4. We evaluate models across 3 families with various sizes. Diamond represents pre-trained only models; circle represents post-trained `Qwen` models; square represents post-trained `Olmo` models; triangle represents post-trained `Llama` models.

Table 5: Parameter grids used for decoding sweeps. For each parameter setting, we sample 1000 generations to report semantic and lexical diversity.

Decoding strategy	Parameter	Values swept
Temperature sampling	T	$\{0.01, 0.3, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.5, 3.0\}$
top-k	k	$\{10, 20, 50, 80, 100, 500\}$
top-p	p	$\{0.1, 0.5, 0.7, 0.9, 0.95\}$
min-p	$p_{\min}$	$\{0.01, 0.1, 0.5, 0.9\}$

as

$$\text{Embedding Diversity} = \frac{2}{n(n-1)} \sum_{i<j} (1 - \cos(\mathbf{e}_i, \mathbf{e}_j)).$$

Higher values indicate higher semantic diversity.

Self-BLEU (Zhu et al., 2018) Given generations $\{y_1, \ldots, y_n\}$ for a single task, we compute

$$\text{Self-BLEU} = \frac{1}{n} \sum_{i=1}^{n} \text{BLEU}(y_i, \{y_j\}_{j \neq i}) \in [0, 1].$$

where BLEU (Papineni et al., 2002) measures the n-gram overlap between a candidate output and a set of reference outputs. In Self-BLEU, lower values indicate higher lexical diversity. We use BLEU-4 in our experiments.

C Shape Calibration Trade-offs

C.1 Random Number Generation

Random Number Generation Prompts

Unconstrained Generation

```
Generate {n} random integers between {start} and {end} (inclusive).
Your response should be {n} integers separated by commas and no white
spaces. Answer:
```

Constrained Generation

```
Generate {n} random integers between {start} and {end} (inclusive)
that sum to at most {sum_value}. Your response should be {n} integers
separated by commas and no white spaces. Answer:
```

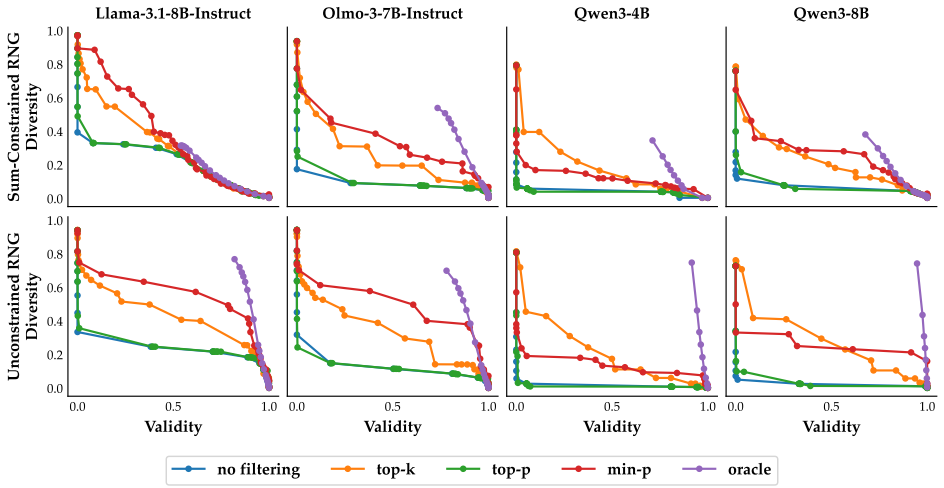


Figure 9: Validity–Diversity trade-off frontiers for constrained and unconstrained random number generation across 4 model families.

Figure 9 shows the validity–diversity trade-off frontiers across all model families. The cutoff oracle helps disentangle the effects of shape and order calibration. In the unconstrained setting, the oracle often approaches the ideal point $(1.0, 1.0)$, suggesting that the main gap stems from standard top-token filters' inability to infer the correct support size. In contrast, in the sum-constrained setting, the oracle itself remains separated from $(1.0, 1.0)$, indicating stronger order miscalibration. The additional gap between standard top-token filtering methods and the oracle reflects shape miscalibration, since fixed top-k, cumulative-mass, and relative-probability thresholds do not reliably recover the correct valid-token boundary. Together, these results show that shape and order miscalibration jointly contribute to the validity–diversity trade-off, with their interaction becoming more pronounced under compositional constraints.

C.2 Random State Generation

Name a Random State Prompt

```
Randomly name a US state with no additional explanation. Answer:
```

Figure 10 shows sequence-level probabilities under different temperatures on the random state generation task. As in Figure 5, the distribution is highly concentrated and heavy-tailed: a small number of valid sequences receive much larger probability than the rest. Increasing temperature flattens the valid-sequence distribution, but it also shifts substantial probability mass into the invalid region before the model approaches a uniform distribution over valid states. Thus, temperature scaling again improves valid diversity only by sacrificing validity.

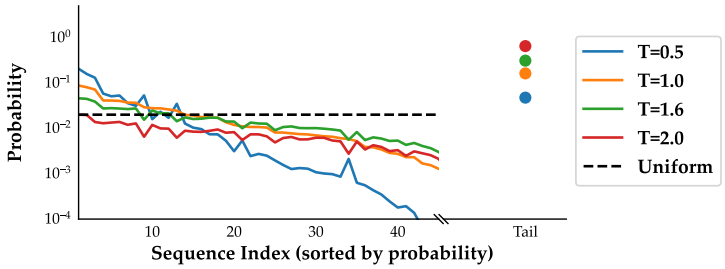


Figure 10: Sequence probability for random state generation task. Sequences are sorted by probability, and probabilities are plotted in log-space. The "tail" section represents the total probability mass of the invalid region.

D Formal Definitions of Calibration

A decoding rule achieves high validity if it assigns high probability only on valid tokens at each time step and it achieves high diversity if it explores many distinct tokens in G rather than concentrating on only a few of them. Therefore, we define order and shape calibration as:

Definition D.1 (Calibration)**.** Given $y_{<t} \in \mathcal{V}^{t-1}$ and an LLM conditional distribution p,

1. **(Order Calibration)** p is order calibrated if for any valid token $v \in \mathcal{V}$ and invalid token $b \in \mathcal{V}$, it assigns a higher probability to the valid token, i.e., $p(v \mid y_{<t}) > p(b \mid y_{<t})$.
2. **(Shape Calibration)** p is shape calibrated if for any token $v \in \mathcal{V}$, it assigns probability mass to v according to the number of valid continuations starting with v, i.e., $p(v \mid y_{<t}) \propto N(y_{<t} \circ v)$.

Shape calibration implies order calibration, since it assigns zero mass to every invalid token and positive mass to every valid one. However, the converse does not hold. Uniformity is one choice of target. For any target distribution π with full support on V, say that p is shape calibrated to π at $y_{<t}$ if $p(\cdot \mid y_{<t}) = \pi(\cdot \mid y_{<t})$, and measure diversity by $\exp[-D_{\mathrm{KL}}(\tilde{p} \,\|\, \pi)]$, which reduces to Definition 3.1 when π is uniform (Equation 2). By the chain rule, p is shape calibrated to π at every valid prefix exactly when $\mathrm{Val}(p) = 1$ and $\tilde{p} = \pi$. The substance of our results is how local deviations from shape calibration compound and which sampling methods cannot undo them. For the fixed-length setting of Appendix E, Lemma E.2 holds with $w(\cdot \mid h) = \pi(\cdot \mid h)$ and diversity $\exp[-D_{\mathrm{KL}}(Q(\cdot \mid V)\|\pi)]$. Theorem 4.2 extends provided the paired valid tokens carry at least γ conditional mass under π at each relevant prefix; the fraction of completion counts must be replaced by this target-mass condition. Theorem 5.2 uses the uniform target distribution.

Our notation of "calibration" is not directly related to LLM confidence calibration literature.

E Analysis of Theorem 4.2

We first retain the multiplicative decomposition of uniform-cutoff precision and support recall as a diagnostic identity. We then prove Theorem 4.2 for weighted, rank-preserving decoding using a common affinity measure.

E.1 Setup and notation

Let $\mathcal{V}$ be a finite vocabulary and let $V \subseteq \mathcal{V}^d$ be the valid sequences of a fixed length $d \geq 1$. For a prefix h, write hv for its concatenation with token v and define

$$N(h) = |\{y \in V : y \text{ starts with } h\}|, \qquad G(h) = \{v \in \mathcal{V} : N(hv) > 0\}.$$

A prefix is valid if $N(h) > 0$. Let U be uniform on V. At every valid prefix of length less than d, its next-token probabilities are

$$w(v \mid h) = \frac{N(hv)}{N(h)}. \tag{1}$$

Thus $U(h) = N(h)/|V|$, $\sum_{v \in G(h)} w(v \mid h) = 1$, and invalid children have target probability zero.

A decoder specifies a normalized distribution $r(\cdot \mid h)$ at every prefix of length less than d, including prefixes having zero probability under the decoder. These kernels induce

$$Q(y) = \prod_{t=1}^{d} r(y_t \mid y_{<t}).$$

For any prefix h, $Q(h)$ denotes the probability that an output begins with h. Whenever $s := Q(V) > 0$, let $P = Q(\cdot \mid V)$. Definition 3.1 gives

$$\mathrm{Val}(Q) = s, \qquad \mathrm{Div}(Q) = \frac{\exp H(P)}{|V|} = \exp[-D_{\mathrm{KL}}(P\|U)]\,. \tag{2}$$

All logarithms are natural, with $0 \log 0 = 0$.

At each prefix, fix a decreasing ordering of the model probabilities $p(\cdot \mid h)$. A decoder is rank preserving if $r(v_i \mid h) \geq r(v_j \mid h)$ whenever $i < j$ in this ordering. This includes positive-temperature sampling followed by a prefix cutoff, with arbitrary prefix-dependent temperatures and cutoffs. Tie handling must agree with the fixed ordering.

E.2 Diagnostic decomposition for uniform cutoffs

For this subsection only, let $S_t(h) \subseteq \mathcal{V}$ be a nonempty retained set and let Q_S sample uniformly from that set at position t. At a valid prefix, define

$$\mathrm{Prec}_t(S; h) = \frac{|S_t(h) \cap G(h)|}{|S_t(h)|}, \qquad \mathrm{Rec}_t(S; h) = \sum_{v \in S_t(h) \cap G(h)} w(v \mid h).$$

The corresponding sequence quantities are

$$\mathrm{Prec}_{\mathrm{seq}}(S) = Q_S(V), \qquad \mathrm{Rec}_{\mathrm{seq}}(S) = \frac{|\{y \in V : y_t \in S_t(y_{<t}) \text{ for every } t\}|}{|V|}.$$

Theorem E.1 (Exact multiplicative decomposition)**.** *Assume $Q_S(V) > 0$. Then*

$$\mathrm{Prec}_{\mathrm{seq}}(S) = \prod_{t=1}^{d} \alpha_t(S), \qquad \mathrm{Rec}_{\mathrm{seq}}(S) = \prod_{t=1}^{d} \beta_t(S),$$

where

$$\alpha_t(S) = \mathbb{E}_{Q_S}[\mathrm{Prec}_t(S; Y_{<t}) \mid Y_j \in G(Y_{<j}) \textit{ for every } j < t]\,,$$
$$\beta_t(S) = \mathbb{E}_U[\mathrm{Rec}_t(S; Y_{<t}) \mid Y_j \in S_j(Y_{<j}) \textit{ for every } j < t]\,.$$

Proof. Let

$$F_t = \{Y_j \in G(Y_{<j}) \text{ for every } j \leq t\}, \qquad E_t = \{Y_j \in S_j(Y_{<j}) \text{ for every } j \leq t\},$$

with $F_0 = E_0$ the full sample space. Since valid sequences have length d, $F_d = \{Y \in V\}$. The assumption $Q_S(V) > 0$ ensures $Q_S(F_t) > 0$ for every t. It also implies that at least one valid sequence is reachable, hence $U(E_d) > 0$ and $U(E_t) > 0$ for every t. All conditional probabilities below are therefore defined.

The chain rule gives

$$Q_S(V) = \prod_{t=1}^{d} Q_S(F_t \mid F_{t-1}).$$

Conditioned on F_{t-1} and its prefix, the probability of choosing a valid child under uniform sampling from S_t is $\mathrm{Prec}_t(S; Y_{<t})$. Taking expectations identifies each factor with $\alpha_t(S)$.

Similarly,

$$\mathrm{Rec}_{\mathrm{seq}}(S) = U(E_d) = \prod_{t=1}^{d} U(E_t \mid E_{t-1}).$$

Under U, the next-token law given a prefix is $w(\cdot \mid Y_{<t})$. The event E_{t-1} is determined by that prefix, so the conditional probability of retaining the next token is $\mathrm{Rec}_t(S; Y_{<t})$. Taking expectations identifies the factors with $\beta_t(S)$. □

If no valid sequence is reachable, both sequence quantities are zero; the conditional decomposition above is asserted only in the positive case. The two products average under different prefix distributions. A trade-off observed at individual prefixes does not by itself imply a trade-off between α_t and β_t. Moreover, count precision is a validity probability specifically for uniform sampling within the retained sets. The following argument instead bounds entropy-based diversity for the weighted decoder Q defined above.

E.3 A LOCAL-TO-GLOBAL AFFINITY BOUND

Lemma E.2 (Affinity compounding)**.** *At each nonterminal valid prefix, define*

$$a(h) = \sum_{g \in G(h)} \sqrt{r(g \mid h) w(g \mid h)}.$$

Suppose that, for each $t \in \{1, \dots, d\}$, a number $\rho_t \in [0, 1]$ satisfies $a(h) \le \rho_t$ at every valid prefix of length $t - 1$. If $s = Q(V) > 0$, then

$$\mathrm{Val}(Q)\,\mathrm{Div}(Q) \le \prod_{t=1}^{d} \rho_t^2. \tag{3}$$

Proof. By Cauchy–Schwarz,

$$a(h)^2 \le \left(\sum_{g \in G(h)} r(g \mid h) \right) \left(\sum_{g \in G(h)} w(g \mid h) \right) \le 1.$$

For $0 \le t \le d$, let

$$A_t = \sum_{\substack{h \in \mathcal{V}^t \\ N(h) > 0}} \sqrt{Q(h) U(h)}.$$

The empty prefix has probability one under both laws, so $A_0 = 1$. Expanding valid prefixes into their children gives

$$\begin{aligned} A_t &= \sum_{\substack{|h| = t-1 \\ N(h) > 0}} \sqrt{Q(h) U(h)} \sum_{g \in G(h)} \sqrt{r(g \mid h) w(g \mid h)} \\ &= \sum_{\substack{|h| = t-1 \\ N(h) > 0}} \sqrt{Q(h) U(h)}\, a(h) \le \rho_t A_{t-1}. \end{aligned}$$

Zero-Q prefixes contribute zero, so no conditioning on a null event is needed. Induction yields

$$A_d = \sum_{y \in V} \sqrt{Q(y) U(y)} \le \prod_{t=1}^{d} \rho_t.$$

For $P = Q(\cdot \mid V)$, its affinity with U is

$$B := \sum_{y \in V} \sqrt{P(y) U(y)} = \frac{A_d}{\sqrt{s}}.$$

Jensen's inequality gives

$$\begin{aligned} \log B &= \log \mathbb{E}_{Y \sim P} \left[\sqrt{\frac{U(Y)}{P(Y)}} \right] \\ &\ge \mathbb{E}_{Y \sim P} \left[\frac{1}{2} \log \frac{U(Y)}{P(Y)} \right] = -\frac{1}{2} D_{\mathrm{KL}}(P \| U). \end{aligned}$$

The expectations are over the support of P, where the ratios are defined. Using equation 2,

$$\mathrm{Div}(Q) \le B^2 = \frac{A_d^2}{s} \le \frac{\prod_{t=1}^{d} \rho_t^2}{s}.$$

Multiplying by $s = \mathrm{Val}(Q)$ proves the lemma. The KL–affinity inequality is also the standard relation $D_1 \geq D_{1/2}$ for Rényi divergence (Van Erven & Harremos, 2014). □

Lemma E.3. *Under the premise of Theorem 4.2, at every valid prefix h at one of the m positions, there are distinct valid tokens $g_1, \ldots, g_n$ and distinct invalid tokens $b_1, \ldots, b_n$ such that each b_j precedes g_j in the fixed ordering and $\sum_j w(g_j \mid h) \geq \gamma$.*

Proof. Cutoffs follow the fixed ordering that defines rank preservation; write S_k for the top-k tokens. Let k_0 be the smallest k with $\mathrm{Rec}_t(S_k; h) > 1 - \gamma$. It exists because the full vocabulary has recall one, and the token at rank k_0 is valid because recall increases only at valid tokens. Let $g_1, \ldots, g_n$ be the valid tokens at rank k_0 or below, in rank order. They carry $1 - \mathrm{Rec}_t(S_{k_0-1}; h) \geq \gamma$ of the valid continuations.

Let b_j be the j-th invalid token in rank order. If g_j has rank $k \geq k_0$, then $\mathrm{Rec}_t(S_k; h) > 1 - \gamma$, so $\mathrm{Prec}_t(S_k; h) \leq 1/2$. Since S_k contains at least j valid tokens, it also contains at least j invalid tokens, all of which precede g_j. Hence b_j precedes g_j. □

E.4 Proof of Theorem 4.2

Proof. Let $I \subseteq \{1, \ldots, d\}$ be the m positions in the theorem. At every valid prefix h at a position in I, take the pairs from Lemma E.3. Write

$$J(h) = \{g_1, \ldots, g_{k(h)}\}, \qquad \gamma(h) = \sum_{g \in J(h)} w(g \mid h) \geq \gamma > 0.$$

Rank preservation implies $r(b_j \mid h) \geq r(g_j \mid h)$ for each pair. Since the invalid witnesses are distinct and disjoint from $G(h)$,

$$\begin{aligned} \sum_{g \in G(h)} r(g \mid h) + \sum_{g \in J(h)} r(g \mid h) &\leq \sum_{g \in G(h)} r(g \mid h) + \sum_{j=1}^{k(h)} r(b_j \mid h) \\ &\leq 1. \end{aligned} \tag{4}$$

Set $c_g = 2$ for $g \in J(h)$ and $c_g = 1$ for the remaining valid tokens. Weighted Cauchy–Schwarz yields

$$\begin{aligned} a(h)^2 &= \left(\sum_{g \in G(h)} \sqrt{c_g r(g \mid h)} \sqrt{\frac{w(g \mid h)}{c_g}} \right)^2 \\ &\leq \left(\sum_{g \in G(h)} c_g r(g \mid h) \right) \left(\sum_{g \in G(h)} \frac{w(g \mid h)}{c_g} \right) \\ &\leq 1 - \frac{\gamma(h)}{2} \leq 1 - \frac{\gamma}{2}. \end{aligned}$$

The first factor is at most one by equation 4; the second is $1 - \gamma(h)/2$. At all other positions, $a(h) \leq 1$. Apply Lemma E.2 with $\rho_t = \sqrt{1 - \gamma/2}$ for $t \in I$ and $\rho_t = 1$ otherwise. This gives

$$\mathrm{Val}(Q)\,\mathrm{Div}(Q) \leq (1 - \gamma/2)^m.$$

In particular, if $\mathrm{Val}(Q) \geq 1 - \epsilon$ with $0 \leq \epsilon < 1$, then

$$\mathrm{Div}(Q) \leq \min\left\{1, \frac{(1 - \gamma/2)^m}{1 - \epsilon}\right\} \leq \min\left\{1, \frac{e^{-\gamma m/2}}{1 - \epsilon}\right\},$$

where the last step uses $1 - x \leq e^{-x}$. □

Scope of the conclusion. The theorem bounds entropy-based diversity. Retaining a large set of valid sequences does not ensure high diversity if most receive negligible probability. The obstruction $\gamma(h)$ measures matched valid continuation mass, and each invalid witness is used only once. The assumption must hold at every valid prefix at the specified common positions. Completion-based labels or a few sampled precision–recall curves do not establish this uniform premise; they provide local diagnostics of rank interleaving.

F Analysis of Theorem 5.2

We prove the finite-vocabulary shape bound and then extend the analysis to geometric parameters that vary across valid prefixes. The first result gives an explicit validity-dependent rate; the extension uses the affinity argument from Lemma E.2.

F.1 Finite geometric model and entropy notation

Let $M = |\mathcal{V}| \geq 2$ and let $\varnothing \neq V \subseteq \mathcal{V}^d$ for a fixed finite length $d \geq 1$. Assume invariant valid branching (Assumption 5.1): every valid prefix at position t has exactly v_t valid next tokens. Throughout this section, assume

$$1 \leq v_t \leq \lfloor M/2 \rfloor, \qquad m = |\{t \in \{1, \dots, d\} : v_t \geq 2\}|.$$

In particular, $|V| = \prod_{t=1}^{d} v_t$.

Definition F.1 (Finite geometric ranked model)**.** Fix $T > 0$. At every valid prefix h of length $t-1$, the valid next tokens occupy exactly ranks $0, \dots, v_t - 1$, and the ranked next-token law is

$$p_t^{(T)}(i \mid h) = \frac{(1-q_t)q_t^i}{1-q_t^M}, \qquad i = 0, \dots, M-1, \tag{5}$$
$$q_t = e^{-\lambda_t/T} \in (0,1), \qquad \lambda_t > 0.$$

The parameter q_t depends on the position but not on the valid prefix. Token identities at particular ranks may vary with the prefix. Let $p^{(T)}$ be the induced law on length-d sequences. Kernels after invalid prefixes may be arbitrary normalized distributions, since those prefixes cannot lead to a sequence in V.

For $v \geq 1$ and $a \geq 0$, define

$$p_{v,a}(i) = \frac{e^{-ai/v}}{\sum_{j=0}^{v-1} e^{-aj/v}}, \qquad i = 0, \dots, v-1,$$

and write

$$H_v(a) = H(p_{v,a}), \qquad \Delta_v(a) = \ln v - H_v(a).$$

Thus $\Delta_1(a) = 0$ and $\Delta_v(0) = 0$. The normalized sharpness at position t is

$$z_t = -v_t \ln q_t = \frac{\lambda_t v_t}{T}.$$

All entropies and logarithms in this analysis use natural units. For $Y \sim p^{(T)}$, the diversity from Definition 3.1 is

$$\mathrm{Div}(p^{(T)}) = \frac{\exp H(Y \mid Y \in V)}{|V|}.$$

Lemma F.2 (Entropy loss increases with sharpness)**.** *For every $v \geq 1$, $\Delta_v(a)$ is nondecreasing in $a \geq 0$. Moreover, $\Delta_v(a) > 0$ whenever $v \geq 2$ and $a > 0$.*

Proof. Put $\theta = a/v$ and $Z(\theta) = \sum_{i=0}^{v-1} e^{-\theta i}$. For $p_\theta(i) = e^{-\theta i}/Z(\theta)$,

$$H_v(a) = \ln Z(\theta) + \theta \mathbb{E}_{p_\theta}[i].$$

Differentiation gives

$$\frac{d}{d\theta} H_v(v\theta) = \theta \frac{d}{d\theta} \mathbb{E}_{p_\theta}[i] = -\theta \operatorname{Var}_{p_\theta}(i) \leq 0.$$

Hence Δ_v is nondecreasing. When $v \geq 2$ and $a > 0$, $p_{v,a}$ is nonuniform, so $H_v(a) < \ln v$. □

F.2 Validity–diversity trade-off

Theorem F.3 (Finite geometric validity–diversity trade-off)**.** *Assume the finite geometric ranked model and invariant valid branching as above. Let $0 < \epsilon < 1/2$ and define*

$$L_\epsilon = \ln \frac{1-\epsilon}{\epsilon} > 0.$$

If $\mathrm{Val}(p^{(T)}) \geq 1-\epsilon$*, then*

$$\mathrm{Div}(p^{(T)}) \leq \exp\left(-\sum_{t=1}^{d} \Delta_{v_t}(L_\epsilon)\right). \tag{6}$$

Proof. At every valid prefix at position t, the probability of choosing a valid next token is

$$a_t = \sum_{i=0}^{v_t-1} \frac{(1-q_t)q_t^i}{1-q_t^M} = \frac{1-q_t^{v_t}}{1-q_t^M}. \tag{7}$$

This quantity is independent of the prefix. A sequence lies in V precisely when every chosen token preserves a valid prefix. The chain rule therefore gives

$$\mathrm{Val}(p^{(T)}) = \prod_{t=1}^{d} a_t. \tag{8}$$

Since $0 < a_t \leq 1$, the premise implies $a_t \geq 1-\epsilon$ for every t. Furthermore, $M \geq 2v_t$ implies

$$a_t = \frac{1-q_t^{v_t}}{1-q_t^M} \leq \frac{1-q_t^{v_t}}{1-q_t^{2v_t}} = \frac{1}{1+q_t^{v_t}}.$$

Consequently,

$$q_t^{v_t} \leq \frac{\epsilon}{1-\epsilon}, \qquad z_t = -v_t \ln q_t \geq L_\epsilon. \tag{9}$$

It remains to justify conditioning on complete validity. From any valid child at position t, the probability of a valid completion is $\prod_{u=t+1}^{d} a_u$. This is independent of the chosen valid child, including at $t=d$, where the empty product is one. Bayes' rule therefore gives, for every valid prefix h and $0 \leq i < v_t$,

$$\begin{aligned} p^{(T)}(\mathrm{rank}(Y_t) = i \mid Y_{<t} = h,\, Y \in V) &= \frac{p_t^{(T)}(i \mid h) \prod_{u>t} a_u}{a_t \prod_{u>t} a_u} \\ &= \frac{q_t^i}{\sum_{j=0}^{v_t-1} q_t^j} = p_{v_t,z_t}(i). \end{aligned}$$

Permutation of token identities does not change entropy. Applying the entropy chain rule under $p^{(T)}(\cdot \mid V)$ yields

$$H(Y \mid Y \in V) = \sum_{t=1}^{d} H_{v_t}(z_t).$$

Together with $|V| = \prod_{t=1}^{d} v_t$, this gives the exact identity

$$\mathrm{Div}(p^{(T)}) = \exp\left(-\sum_{t=1}^{d} \Delta_{v_t}(z_t)\right). \tag{10}$$

Apply equation 9 and Lemma F.2 term by term to obtain equation 6. □

Lemma F.4 (Entropy loss increases with the valid branching count)**.** *For every fixed $L > 0$, $\Delta_v(L)$ is strictly increasing in the positive integer v. In particular, $\Delta_v(L) \geq \Delta_2(L)$ for all $v \geq 2$.*

Proof. For $\theta > 0$, the normalizing constant of the geometric law on $\{0, \ldots, v-1\}$ is

$$Z_v(\theta) = \sum_{i=0}^{v-1} e^{-\theta i} = \frac{1-e^{-v\theta}}{1-e^{-\theta}}.$$

Its expected rank is

$$-\frac{d}{d\theta} \ln Z_v(\theta) = \frac{1}{e^\theta - 1} - \frac{v}{e^{v\theta}-1}.$$

Using $H_v(L) = \ln Z_v(L/v) + (L/v)\mathbb{E}[i]$, we obtain

$$\Delta_v(L) = \ln\left(\frac{v(1-e^{-L/v})}{1-e^{-L}}\right) - \frac{L/v}{e^{L/v}-1} + \frac{L}{e^L-1}. \tag{11}$$

Define, for $x > 0$,

$$g(x) = \ln\left(\frac{1 - e^{-x}}{x}\right) - \frac{x}{e^x - 1}.$$

Rearranging equation 11 gives

$$\Delta_v(L) = C(L) + g(L/v), \qquad C(L) = \ln L - \ln(1 - e^{-L}) + \frac{L}{e^L - 1}.$$

Differentiating g gives

$$g'(x) = -\frac{1}{x} + \frac{xe^x}{(e^x - 1)^2} = -\frac{1}{x} + \frac{x}{4\sinh^2(x/2)} < 0,$$

where the strict inequality follows from $2\sinh(x/2) > x$ for $x > 0$. Thus g is strictly decreasing. As the positive integer v increases, L/v strictly decreases, so $g(L/v)$ and hence $\Delta_v(L)$ strictly increase. □

Corollary F.5 (Exponential loss in branching length)**.** *For $0 < \epsilon < 1/2$, define*

$$\begin{aligned} c(\epsilon) &:= \Delta_2\left(\ln\frac{1-\epsilon}{\epsilon}\right) \\ &= \ln 2 - h_2\left(\frac{\sqrt{1-\epsilon}}{\sqrt{1-\epsilon} + \sqrt{\epsilon}}\right), \end{aligned} \tag{12}$$

where $h_2(a) = -a\ln a - (1-a)\ln(1-a)$ is binary entropy in natural units. Under Theorem F.3,

$$\mathrm{Div}(p^{(T)}) \le e^{-mc(\epsilon)}, \qquad c(\epsilon) > 0, \qquad \lim_{\epsilon \downarrow 0} c(\epsilon) = \ln 2.$$

The rate $c(\epsilon)$ is independent of the vocabulary size M.

Proof. Write $L_\epsilon = \ln((1-\epsilon)/\epsilon) > 0$. The explicit expression in equation 12 follows from $p_{2,L_\epsilon}(0) = \sqrt{1-\epsilon}/(\sqrt{1-\epsilon} + \sqrt{\epsilon})$. Positions with $v_t = 1$ contribute zero entropy loss. At each of the m branching positions, Lemma F.4 gives

$$\Delta_{v_t}(L_\epsilon) \ge \Delta_2(L_\epsilon) = c(\epsilon).$$

Substituting into equation 6 proves the exponential bound. The positivity follows from Lemma F.2. As $\epsilon \to 0$, $L_\epsilon \to \infty$, and p_{2,L_ϵ} concentrates at rank zero. Its entropy tends to zero, so $c(\epsilon) = \ln 2 - H_2(L_\epsilon) \to \ln 2$. This proves Theorem 5.2. □

Interpretation and parameter regime. High sequence validity requires every local validity factor to be close to one. In the finite geometric model, this forces positive sharpness across each valid-token interval. Conditioning on validity leaves a nonuniform distribution over valid choices, and the resulting entropy losses add over positions. Exponentiating entropy converts this sum into multiplicative diversity loss. The bound applies for $0 < \epsilon < 1/2$; its rate tends to zero as $\epsilon \to 1/2$. The rate $c(\epsilon)$ does not depend on the vocabulary size. The next result removes prefix invariance of the geometric parameter, while retaining invariant valid branching.

Heavier tails. The proof uses the invalid ranks only through the bound $a_t \le 1/(1 + q_t^{v_t})$. Suppose the valid-rank logits satisfy $\ell_i = \ell_0 - \lambda_t i$ for $0 \le i < v_t$, but the invalid ranks $i \ge v_t$ have arbitrary logits $\ell_i \ge \ell_0 - \lambda_t i$, that is, on or above the linear extrapolation of the head, depending on the position but not on the prefix. At any temperature the weight of rank i is then at least $e^{\ell_0/T} q_t^i$, so ranks $v_t, \ldots, 2v_t - 1$ alone carry at least $q_t^{v_t}$ times the valid mass, and $a_t \le 1/(1 + q_t^{v_t})$ still holds. The remaining steps use only the valid ranks and the prefix independence of a_t, so Theorem F.3 and Corollary F.5 hold unchanged. A logarithmic tail $\ell_i = A + B\log(i + C)$ with $B < 0$ and $i + C > 0$ throughout the tail satisfies this condition if it meets the extrapolated head at the junction and its slope there is at least $-\lambda_t$: convexity then keeps the tail on or above that line.

F.3 Extension to prefix-dependent geometric shape

Theorem F.6 (Geometric shape with heterogeneous prefixes)**.** *Let $M = |\mathcal{V}| \ge 4$. Suppose every valid prefix at position t has v_t valid children, where $1 \le v_t \le \lfloor M/2 \rfloor$, and these children occupy*

the first v_t ranks. At each valid prefix h, let

$$r(i \mid h) = \frac{(1-q(h))q(h)^i}{1-q(h)^M}, \qquad i = 0, \dots, M-1, \qquad 0 < q(h) < 1.$$

The parameter $q(h)$ may vary arbitrarily across valid prefixes. Let Q be the induced sequence law and $m = |\{t : v_t \geq 2\}|$.

There is a universal constant $0 < \kappa_\star < 1$, independent of vocabulary size, such that

$$\mathrm{Val}(Q)\,\mathrm{Div}(Q) \leq \kappa_\star^m.$$

One can take $\kappa_\star \approx 0.847432$, defined in Equation 14 below.

Proof. Let $s = \mathrm{Val}(Q) > 0$. We first bound the local score uniformly over the geometric parameter and vocabulary size. For $2 \leq v \leq \lfloor M/2 \rfloor$ and $0 < q < 1$, define

$$b_{M,v}(q)^2 = \frac{1-q}{v(1-q^M)} \left(\sum_{i=0}^{v-1} q^{i/2} \right)^2, \tag{13}$$

with continuous endpoint values $b_{M,v}(0)^2 = 1/v$ and $b_{M,v}(1)^2 = v/M$. Let

$$\begin{aligned} \kappa_M &:= \max_{2 \leq v \leq \lfloor M/2 \rfloor} \max_{q \in [0,1]} b_{M,v}(q)^2, \\ \kappa_\star &:= \max_{q \in [0,1]} b_{4,2}(q)^2 \approx 0.847432. \end{aligned} \tag{14}$$

Invariant branching implies that every valid child at position t has $\prod_{u>t} v_u$ valid completions. Therefore, if U_V is uniform on V, then $U_V(g \mid h) = 1/v_t$ on every valid child. At a branching position, the local affinity between the geometric kernel and this target is

$$a(h) = \sum_{i=0}^{v_t-1} \sqrt{\frac{r(i \mid h)}{v_t}} = b_{M,v_t}(q(h)).$$

At a nonbranching position, $a(h) \leq 1$.

The expression in equation 13 has the stated continuous limits: as $q \to 0$, the distribution concentrates at rank zero, and as $q \to 1$, it becomes uniform over the vocabulary. Both endpoint values of $b_{M,v}(q)^2$ are at most $1/2$. For $0 < q < 1$, Cauchy–Schwarz and $v < M$ give

$$b_{M,v}(q)^2 \leq \sum_{i=0}^{v-1} \frac{(1-q)q^i}{1-q^M} < 1,$$

because invalid ranks have strictly positive mass. The continuous function $b_{M,v}(q)^2$ attains its maximum on $[0,1]$. Since it is strictly below one everywhere, its maximum is below one. Taking the maximum over finitely many v still gives $0 < \kappa_M < 1$.

We next bound all vocabulary sizes by the same constant. Since $M \geq 2v$, $b_{M,v}(q)^2 \leq b_{2v,v}(q)^2$ for $0 < q < 1$. Set $u = q^{v/2}$ and $x = -\ln u > 0$. Summing the geometric series gives

$$b_{2v,v}(q)^2 = \frac{1}{v} \coth\!\left(\frac{x}{2v}\right) \frac{(1-u)^2}{1-u^4}.$$

For $z > 0$, the function $z \coth z$ is strictly increasing, since

$$\frac{d}{dz}(z \coth z) = \frac{\sinh z \cosh z - z}{\sinh^2 z} > 0.$$

Therefore $v \geq 2$ implies

$$\frac{1}{v} \coth\!\left(\frac{x}{2v}\right) \leq \frac{1}{2} \coth\!\left(\frac{x}{4}\right), \qquad b_{M,v}(q)^2 \leq b_{4,2}(q^{v/2})^2 \leq \kappa_\star.$$

The endpoint values also satisfy this inequality, so $\kappa_M \leq \kappa_\star < 1$. For the numerical value, writing $\tau = \sqrt{q}$ gives $b_{4,2}(q)^2 = (1+\tau)^2/[2(1+\tau^2)(1+\tau^4)]$. Its derivative has the opposite sign to $2\tau^6+3\tau^5+\tau^4+2\tau^3+\tau-1$, a strictly increasing polynomial with one root in $(0,1)$. Evaluating at this

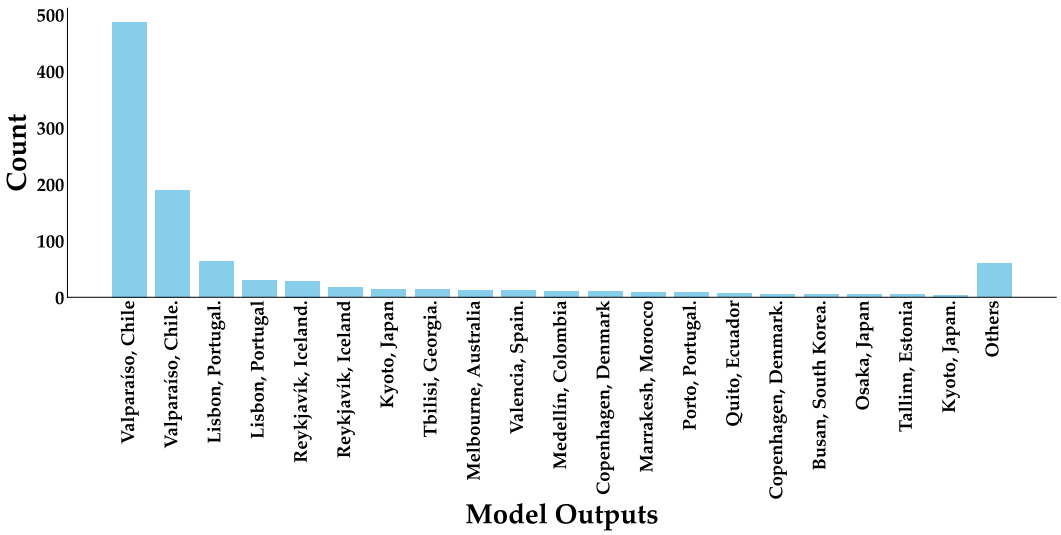


Figure 11: Prompting `GPT-5.5` to randomly name a city in the world. The vast majority of answers, with or without user chat history, collapses to "Valparaíso, Chile." This shows a strong collapse in diversity.

root gives the stated approximation. The constant is attained for $M = 4$, $v = 2$, and $q \approx 0.264217$; it is sharp for the local squared affinity, not necessarily for the final diversity bound.

Apply Lemma E.2 with $\rho_t = \sqrt{\kappa_M}$ at branching positions and $\rho_t = 1$ elsewhere:

$$\mathrm{Val}(Q)\,\mathrm{Div}(Q) \le \kappa_M^m \le \kappa_\star^m.$$

In particular, if $s \ge 1 - \epsilon$ with $0 \le \epsilon < 1$, then

$$\mathrm{Div}(Q) \le \min\left\{1, \frac{e^{-mc_\star}}{1-\epsilon}\right\}, \qquad c_\star = -\ln \kappa_\star \approx 0.165544.$$

□

This extension does not require the entropy factorization in equation 10: conditioning on complete validity can change the relative weights of valid children when future survival depends on the prefix. The affinity bound accounts for that dependence directly. Both shape results assume a geometric ranked law and the stated valid-set structure; they are mechanisms under controlled assumptions rather than claims about every LLM conditional.

G Experiments with Production-Level Models

In Figure 11, we prompted `GPT-5.5` with the prompt "Name a random city in the world", with default thinking level and temperature. The model's response almost always lies in a limited set of a few cities.

H Case Study: Calibration in Coding

We examine the distributional properties of LLMs in a simple coding setting. Although diversity is not inherently the primary objective in coding tasks, prior work has shown that greedy decoding is often suboptimal. Instead, Yue et al. (2025) advocate evaluating models using the *pass@k* metric, which measures whether at least one of k sampled solutions correctly solves the task. This highlights the importance of effective sampling even in domains where a single correct answer suffices.

Figure 12 illustrates sequence-level probability distributions on a coding task from LiveCodeBench (Jain et al., 2025). We observe clear evidence of both order and shape miscalibration. Valid solutions are interspersed with invalid ones throughout the ranked distribution (Karan & Du, 2026), indicating poor order calibration. At the same time, probability mass is unevenly concentrated across valid solutions, with a small subset dominating the distribution, reflecting shape miscalibration.

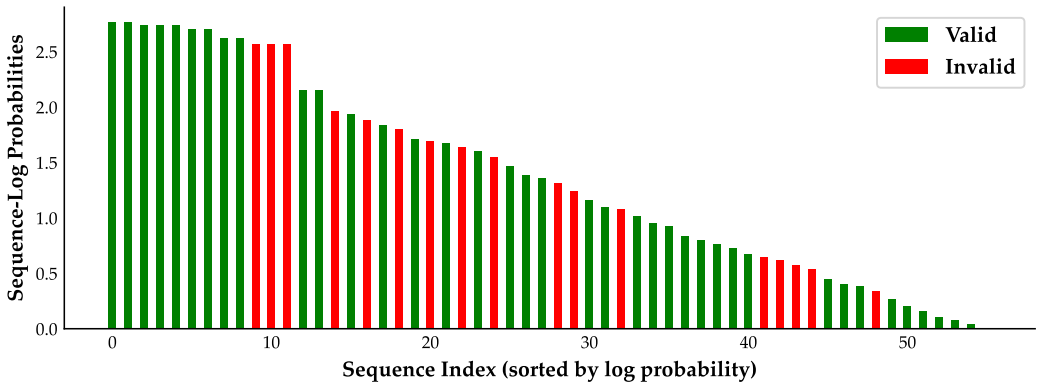


Figure 12: Sequence-level probability distribution for a coding task from LiveCodeBench (Jain et al., 2025), with sequences sorted by log-probability. Valid and invalid solutions are intermixed across the ranking (order miscalibration), and probability mass is concentrated on a small subset of valid solutions (shape miscalibration).

These results show that our framework helps diagnose model failures beyond diversity alone, providing a lens to understand why sampling-based improvements such as higher *pass@k* remain difficult to achieve in practice.

I Empirical Analysis of Logits

To investigate the logits distribution, we fit the logits at each generation conditional to a piecewise model, defined as

$$f(k) = \begin{cases} mk + b, & k \leq c, \\ A + B\log(k + C), & k > c, \end{cases}$$

where k refers to the logit index and $f(k)$ the corresponding logit value. To find c, we sweep over all token indexes. Figures 13, 14, 15 show the curve fit and MSE, R^2 of fitting to various tasks and conditionals. Results show that LLM conditionals are consistently sharp-headed and heavy-tailed.

J Limitations

Our work studies diversity collapse through the lens of validity–diversity calibration. This perspective is useful because it connects output-level diversity failures to local properties of the model's conditional distributions. However, several limitations should be noted.

Controlled tasks are diagnostic rather than exhaustive. A substantial part of our empirical analysis uses controlled random-number generation tasks, where the valid set can be characterized exactly. This allows us to compute validity and diversity without relying on noisy semantic judgments, and makes it possible to isolate the effects of shape and order calibration. However, these tasks are not intended to capture the full complexity of open-ended generation. In domains such as creative writing, scientific ideation, dialogue, or planning, validity is semantic, context-dependent, and often graded rather than binary. Therefore, our controlled experiments should be interpreted as diagnostics for specific distributional mechanisms, not as a complete account of all diversity failures in realistic generation settings.

Token-level validity labels are approximate in open-ended settings. For open-ended tasks, the valid-token set is not directly observable. Our empirical procedure approximates token validity by extending a prefix with a candidate token, greedily decoding a completion, and then evaluating the final sequence with an LLM judge. This provides a practical estimate of whether a token preserves access to a valid continuation, but it is not an exact characterization of the true valid-token set: a token that leads to an invalid greedy continuation may still admit valid continuations under another decoding path.

To assess the reliability of this approximation, we conduct a human validation study comparing judge-based labels against human annotations. This helps quantify the extent to which the LLM judge agrees with human judgments and reduces concern that our precision–recall estimates are

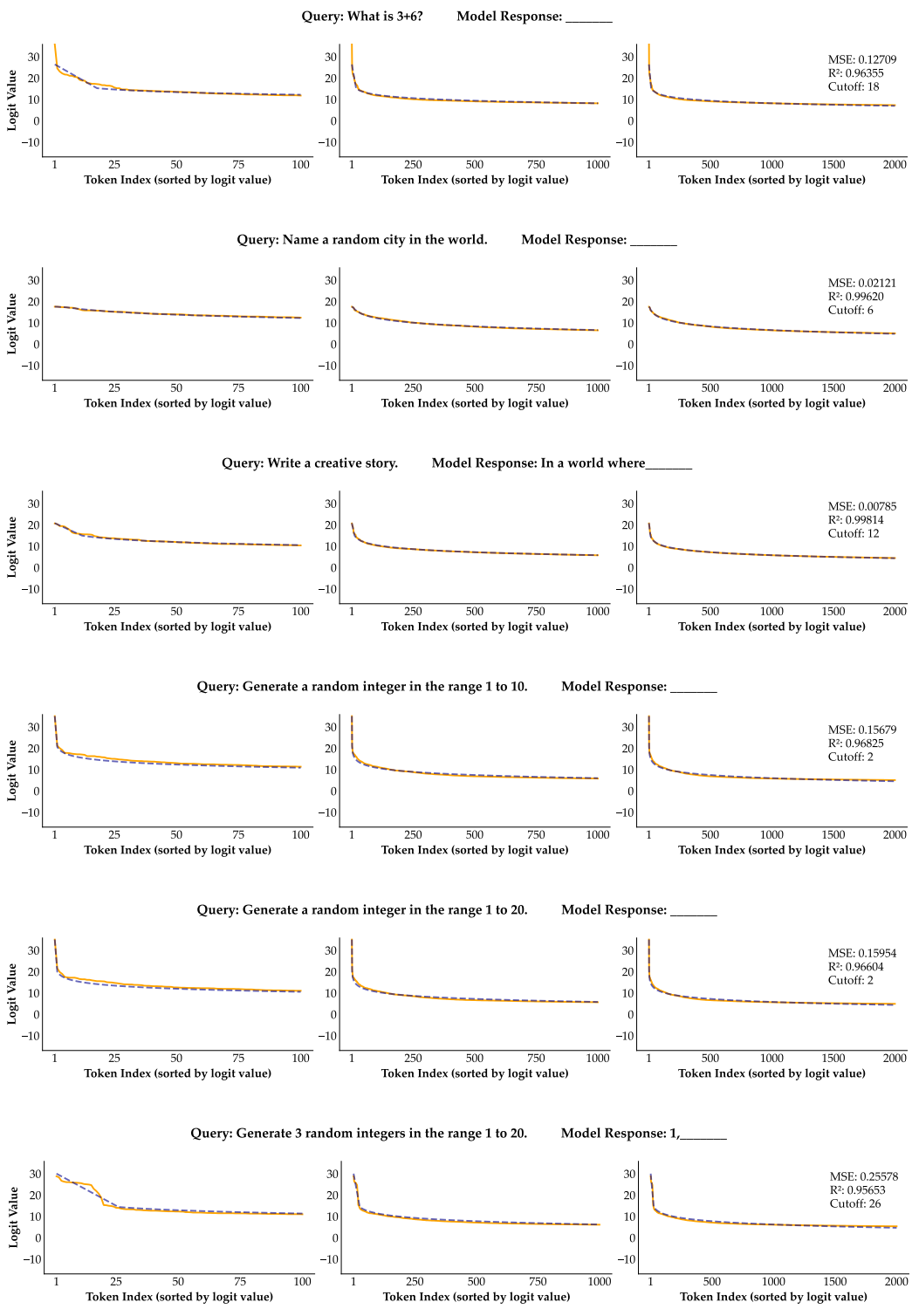


Figure 13: Logit fitting on `Llama-3.1-8B-Instruct`

artifacts of a particular judge model. Nevertheless, the labels remain approximate, since both human and model judgments depend on the task rubric, validity threshold, and the particular continuation used for evaluation. Our controlled random-number experiments avoid this source of noise by using algorithmically known validity.

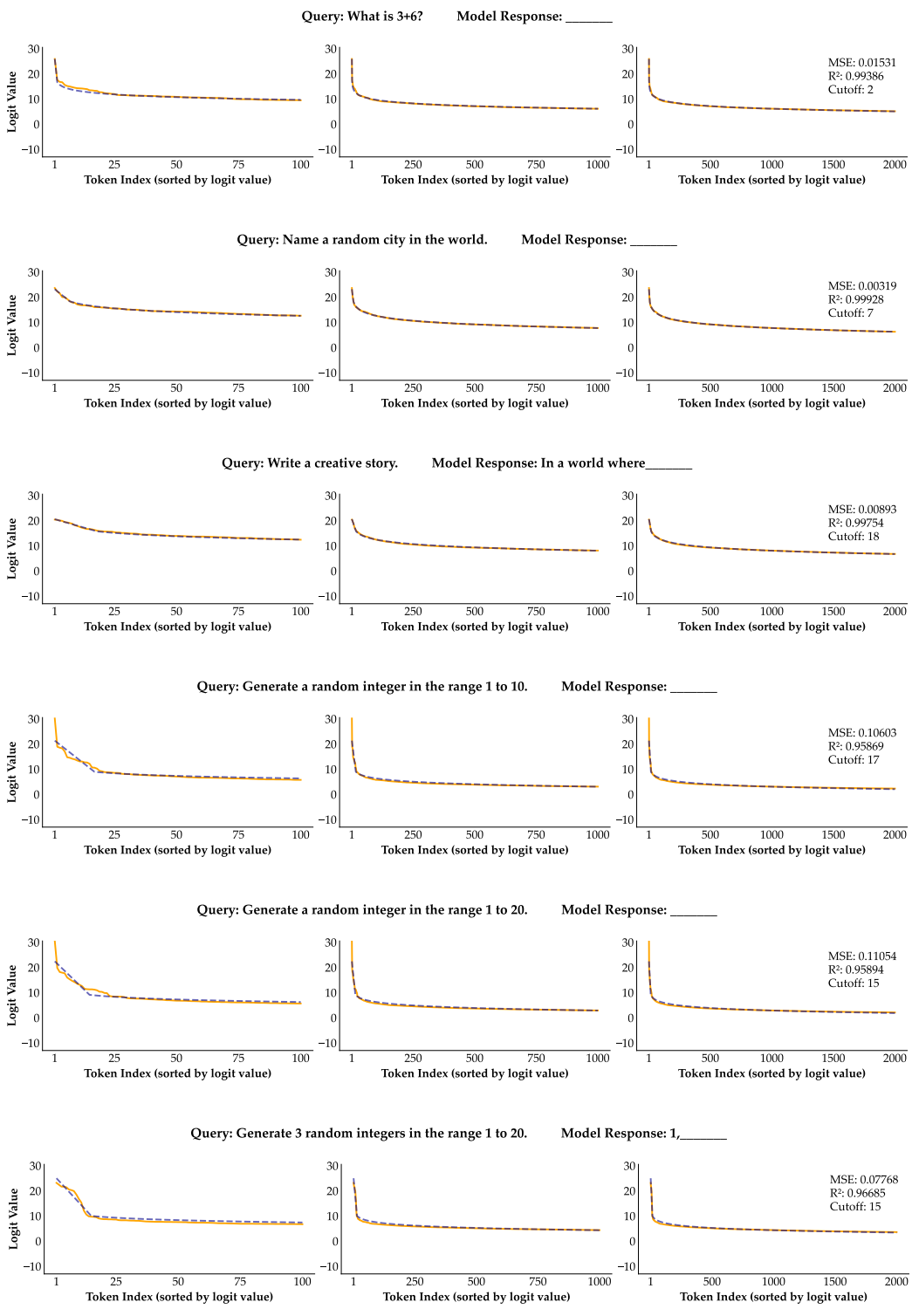


Figure 14: Logit fitting on `Qwen3.5-35B-A3B`

The theoretical models isolate mechanisms under stylized assumptions. Our theoretical results are designed to formalize clean mechanisms rather than to fully model all LLM distributions. In particular, the order theorem requires persistent rank inversions covering a fixed fraction of valid completions at every relevant valid prefix. The shape calibration analysis assumes a normalized finite geometric distribution, invariant valid branching, and at least as many invalid as valid tokens;

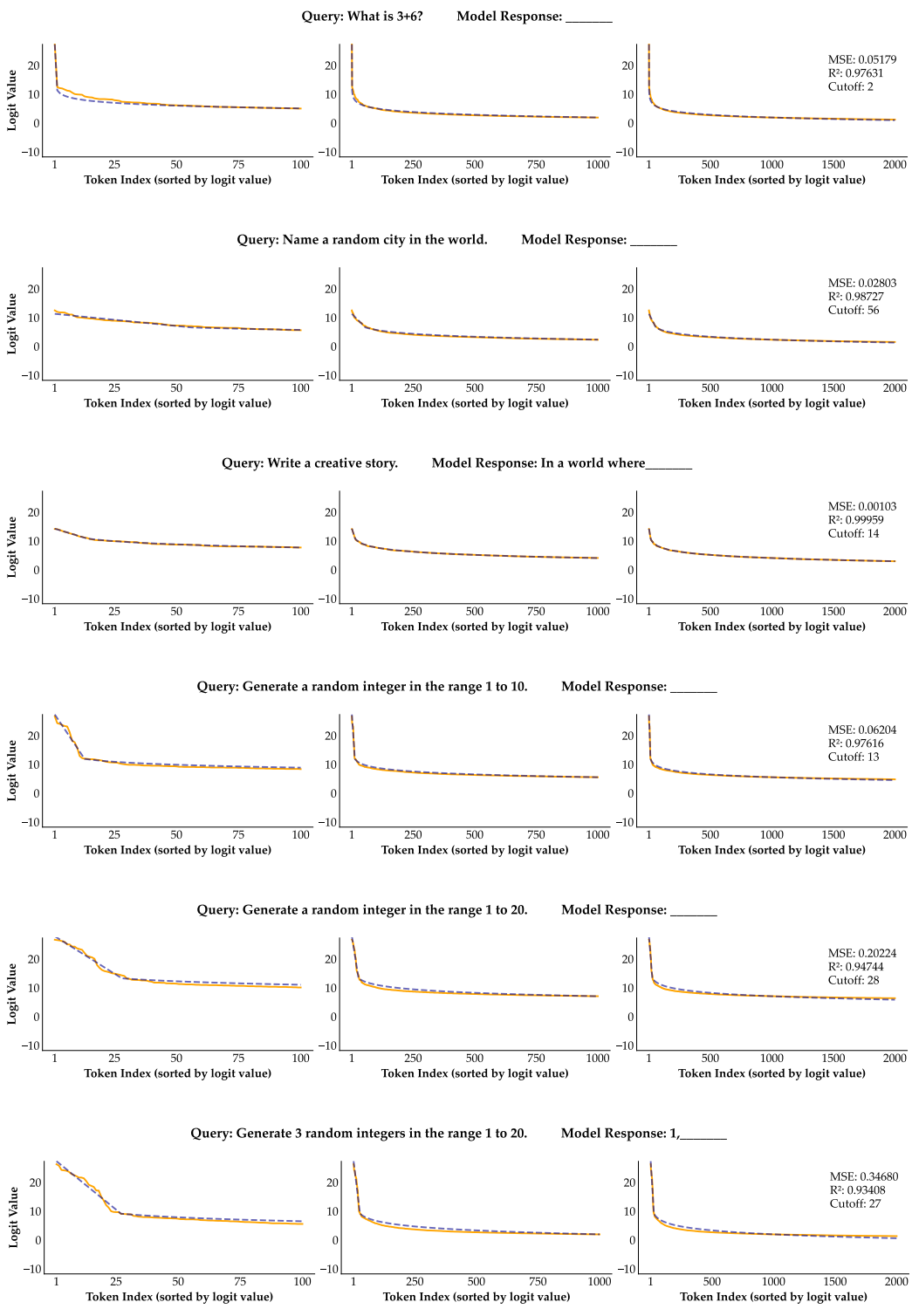


Figure 15: Logit fitting on `Olmo-3-7B-Instruct`

Theorem 5.2 additionally assumes the same ranked distribution across valid prefixes at each position. These assumptions make it possible to show how local sharpness and entropy losses compound across sequence positions. Real LLM conditionals may have heterogeneous branching factors, non-geometric tails, prefix-dependent valid sets, and interactions between syntax, semantics, and instruction-following constraints. The theory should therefore be read as a mechanistic explanation

of why validity–diversity trade-offs can arise, rather than as a literal generative model of all LLM behavior.

Oracle baselines are diagnostic, not deployable. Several experiments use oracle information, such as the ground-truth valid-token set size or exact validity constraints in controlled tasks. These oracle baselines are not meant to be practical decoding methods. Their purpose is to separate failure modes. For example, an oracle-size cutoff tests whether a rank-based method would improve if it knew the correct local support size, while still failing when valid and invalid tokens are interleaved in rank. Thus, oracle performance should be interpreted as evidence about the source of the bottleneck, not as a directly available inference-time algorithm.

Sequence-level experiments are limited by sequence depth. Exact validity–diversity computation becomes expensive as sequence length grows, because the number of possible continuations increases rapidly. This limits the lengths and branching structures that can be exhaustively evaluated in controlled settings. Our experiments therefore emphasize short-to-moderate horizons where exact computation is feasible. The theory predicts that the relevant losses compound with the number of branching positions, but larger-scale empirical validation over longer sequences remains an important direction for future work.

This work diagnoses rather than solves the bottleneck. Our goal is to identify distributional mechanisms that constrain validity and diversity during decoding. We do not propose a new decoding algorithm or training objective that fully resolves these issues. The results suggest that effective solutions may need to improve the model's conditional calibration directly, or use auxiliary validity signals that go beyond probability rank. Designing such calibration-aware training or decoding methods is left for future work.

K Compute Resources

We conduct our experiments on 8 NVIDIA A6000 GPUs. For experiments on commercial models, we use the corresponding official API endpoints. Open-source LLM inference is conducted through the vLLM (Kwon et al., 2023) package. The total computing time required to reproduce all our generations and evaluate the results is around 1 week.

L Societal Impacts

Our work directly addresses the frontiers of LLM inference. In particular, our work aims to understand the validity and diversity of LLM-generated outputs through fine-grained token control. Although we observe no evidence for harmful content in our experiments, the token-level control could potentially lead to instances where sensitive/harmful content is produced. However, our method does not attempt to jailbreak or induce unsafe behavior in any model.

On a high level, work incentivizing LLM generation diversity can produce novel ideas and solutions to societal problems, but also increase risk of biased content. Advancing model diversity and capability while guaranteeing trustworthiness and safety remains a high priority in this and all future works.