TEMPERATURE AND TRUNCATION DO NOT COMMUTE: EXACT CONDITIONS FOR ORDER-INDEPENDENT DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Two systems that report identical sampling settings can sample from different distributions. Settings are reported as a short tuple such as "temperature 0.7, top-p 0.95", while temperature scaling and truncation are applied in sequence to the logit vector and the order of the two steps is fixed by the inference implementation, not by the tuple. Temperature and min-p truncation are order independent under the reparameterization $\alpha \mapsto \alpha^\tau$, a closed-form correction that depends on neither the model nor the vocabulary. Temperature and top-p are not order independent in general, and at fixed $\tau \neq 1$ and fixed $p_0 < 1$ no change of the top-p threshold alone can make the two orders agree for every distribution, so the natural remedy of retuning p_0 is unavailable. Below unit temperature the effect is bounded by $1 - p_0$, and the same bound explains why it is largest at intermediate temperatures and vanishes as τ falls further. Over the complete vocabulary on approximately $290{,}000$ next-token distributions per model, the corrected min-p pairing gives zero total variation distance in all 24 tested combinations, while the uncorrected min-p and top-p pairings reach 0.231 and 0.496 at $\tau = 1.5$ and are zero only at $\tau = 1$. The Transformers library instantiates one of the two orders by default. On GSM8K, swapping the top-p order under a shared random stream changes exact-match accuracy by a statistically significant 1.34 percentage points. Reporting the composition order, or adopting the corrected min-p threshold, makes the reported configuration determine the sampled distribution.

1 INTRODUCTION

Stochastic decoding applies two operators to the logit vector a language model produces. A temperature rescales the logits, a truncation rule keeps only a high-probability subset of the vocabulary (Holtzman et al., 2020; Nguyen et al., 2025), and the next token is drawn from what remains. Papers, model cards, and inference logs summarize these settings as a small tuple, for example "temperature 0.7, top-p 0.95". The procedure contains a choice that the tuple does not record, because the implementation must apply one of the two operations first. Two systems that report the same (τ, p_0) but apply the two operations in different orders can sample from different distributions while stating identical settings.

Much of current language model evaluation is stochastic. Self-consistency estimates (Wang et al., 2023), pass@k code-generation metrics (Chen et al., 2021), sampling-based uncertainty estimates (Manakul et al., 2023; Kuhn et al., 2023), and diversity measurements (Ippolito et al., 2019) all draw many samples under a reported configuration and average over them. If the configuration tuple does not fix the sampling distribution, then two papers, or two deployments of the same model, can report the same settings and still obtain systematically different evaluation numbers, and the difference cannot be traced from the reported settings alone. Decoding hyperparameters are already known to change generated text substantially, and to do so differently across tasks (Wiher et al., 2022; Garces Arias et al., 2025). Reporting only (τ, p_0, α) leaves the composition order implicit, and reading the tuple as a complete specification assumes that temperature and truncation commute.

Recent work answers this interaction by designing a truncation rule that is invariant to temperature (Ding et al., 2026). This paper instead characterizes the operators already in use and measures

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.

what the composition order changes. The theory, proved for arbitrary finite categorical distributions, separates the two truncation rules most often paired with temperature. One of them admits an exact reparameterization of its threshold that makes the order immaterial, and the other admits none. We test this theory on three language models, evaluating both orders at every position. A precision and vocabulary-restriction ablation then compares the order effect with the errors introduced by reduced precision and by a measurement limited to the top 1024 logits. We also execute both orders with the logits processors of the Transformers library. Finally we measure the change in GSM8K exact-match accuracy under a shared-uniform coupling that controls for variation from independent draws.

Our contributions are:

- An exact characterization, with full proofs, of order independence for the four truncation rules in common use. Min-p admits a closed-form reparameterization of its threshold, top-p admits none at fixed τ and fixed $p_0 < 1$ and has an order effect bounded by $1 - p_0$ below unit temperature that accounts for its non-monotonic dependence on τ, and top-k and min-k need no correction.
- A full-vocabulary measurement of both orders on three pretrained language models over the complete WikiText-103 test split, which verifies the min-p identity at every one of roughly $870{,}000$ evaluated positions, together with a precision and vocabulary-restriction ablation that quantifies the numerical error a top-k-limited measurement would introduce.
- An implementation study showing that the Transformers logits processors instantiate a specific composition order by default, a choice that the temperature and threshold tuple does not encode.
- A sequence-level study on GSM8K, using a shared-uniform coupling, showing that the composition order changes exact-match accuracy by a statistically significant 1.34 percentage points, together with an entropy-stratified analysis of where the order effect concentrates.

2 Related Work

Stochastic decoding truncates the next-token distribution before a token is drawn. Top-k truncation keeps a fixed number of highest-probability tokens (Fan et al., 2018), and nucleus (top-p) sampling keeps the shortest prefix of the sorted distribution whose cumulative mass reaches a threshold (Holtzman et al., 2020). Min-p sampling instead keeps tokens whose probability exceeds a fixed fraction of the largest probability, so the size of the retained set follows the confidence of the model at each step (Nguyen et al., 2025). Nadeem et al. (2020) analyze top-k, nucleus, and temperature sampling as transformations of the next-token distribution and identify properties they share, such as entropy reduction and order preservation. More recent rules make the threshold adaptive: Mirostat adjusts a rank-based cut to control perplexity (Basu et al., 2021), η-sampling sets an entropy-dependent threshold and interprets truncation as desmoothing (Hewitt et al., 2022), locally typical sampling keeps tokens whose surprisal is close to the conditional entropy (Meister et al., 2023), and Top-H states truncation as an entropy-constrained mass-maximization problem (Baghaei Potraghloo et al., 2025). Finlayson et al. (2024) study why threshold truncation helps and when it removes tokens that carry true probability mass.

Temperature scaling raises every probability to the power $1/\tau$ and renormalizes, a transformation also used for post-hoc calibration (Guo et al., 2017). Empirical studies report that decoding hyperparameters change generation quality substantially and that their effect is task dependent (Wiher et al., 2022; Garces Arias et al., 2025), and recent work reports mixed empirical findings for min-p sampling (Schaeffer et al., 2025). Closest to the present paper, Ding et al. (2026) observe that probability-space truncation rules such as top-k, top-p, and min-p are sensitive to temperature scaling, and respond by proposing min-k, a logit-space rule that they prove to be temperature invariant.

This paper studies composition order directly. Ding et al. (2026) address temperature sensitivity by introducing a new operator that removes the interaction by construction. We instead characterize the operators already in use, giving necessary and sufficient conditions for the two orders of an existing pair to produce the same distribution, an exact threshold reparameterization for min-p, an impossibility result for top-p, and a measurement of its effect at the distribution level and at the benchmark level on three language models. This bears on evaluation because many procedures aggregate over

repeated samples and therefore depend on the exact sampling distribution. Self-consistency selects the most frequent answer among sampled reasoning paths (Wang et al., 2023), pass@k aggregates independent samples per problem (Chen et al., 2021), and sampling-based uncertainty methods measure agreement among repeated generations (Manakul et al., 2023; Kuhn et al., 2023). Nucleus size has also been reported to affect how much training text a model reproduces (Borec et al., 2024).

3 Background and Notation

Let V denote the model vocabulary with $|V| = n < \infty$. At a decoding position, the model produces a logit vector $\ell \in \mathbb{R}^n$. The base categorical distribution $p \in \Delta^{n-1}$, the probability simplex over V, is obtained by the softmax

$$p_i = \frac{\exp(\ell_i)}{\sum_{j=1}^{n} \exp(\ell_j)}. \tag{1}$$

We sort indices so that $p_1 \geq p_2 \geq \cdots \geq p_n$, breaking ties deterministically by ascending token identifier. This ordering defines the nucleus prefixes below, and temperature scaling preserves it.

Definition 1 (Temperature operator). *For $\tau \in \mathbb{R}_{>0}$, the temperature operator $Q_\tau : \Delta^{n-1} \to \Delta^{n-1}$ is*

$$Q_\tau(p)_i = \frac{p_i^{1/\tau}}{\sum_{j=1}^{n} p_j^{1/\tau}}, \tag{2}$$

with the convention $0^{1/\tau} = 0$.

Definition 2 (Nucleus, or top-p, operator (Holtzman et al., 2020)). *For $p_0 \in (0, 1]$, let $m(p) = \min\{k \in \{1, \ldots, n\} : \sum_{i \leq k} p_i \geq p_0\}$ be the minimal admissible prefix index. The nucleus operator N_{p_0} is*

$$N_{p_0}(p)_i = \begin{cases} p_i / \sum_{j \leq m(p)} p_j & i \leq m(p) \\ 0 & \textit{otherwise.} \end{cases} \tag{3}$$

Definition 3 (Min-p operator (Nguyen et al., 2025)). *For $\alpha \in (0, 1]$, let $S_\alpha(p) = \{i : p_i \geq \alpha \max_j p_j\}$. The min-$p$ operator M_α is*

$$M_\alpha(p)_i = \begin{cases} p_i / \sum_{j \in S_\alpha(p)} p_j & i \in S_\alpha(p) \\ 0 & \textit{otherwise.} \end{cases} \tag{4}$$

We measure disagreement between two distributions $p, q \in \Delta^{n-1}$ obtained from the same base logits under different composition orders using total variation (TV) distance,

$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{i=1}^{n} |p_i - q_i| \in [0, 1]. \tag{5}$$

We use total variation distance because, at a decoding position where the two orders share the same conditioning context, the probability that they select different tokens is at least $\mathrm{TV}(p, q)$ under any coupling. The per-position measurements therefore lower-bound the probability that the two orders diverge at the first step at which they are compared.

We write order A for the composition that applies temperature first and order B for the composition that applies truncation first. For the top-p pair we also track the nucleus-index difference $|m_A - m_B|$ between the two orders, the retained base-mass difference between the two truncated supports, and the ratio of the tempered-nucleus support size to the base-nucleus support size. These quantities separate how many tokens the two orders retain differently from how much probability mass that difference carries, and the two need not vary proportionally: a nucleus that grows by thousands of ranks changes the sampled distribution little when the added tokens are individually improbable. Because both truncated supports under the top-p pair are prefixes of the same sorted order, one support always contains the other.

Shannon entropy $H(p) = -\sum_i p_i \log p_i$ is used to stratify all of these quantities by how peaked the underlying next-token distribution is. The two orders can differ only where temperature scaling

moves the truncation boundary past a token, which requires several tokens near the boundary to carry comparable probability, so the effect concentrates at positions where the distribution is flat.

For sequence-level analysis we generate two completions of length up to T from the same context under two different composition orders, coupling the randomness with a single stream of independent uniform variates $U_1, \ldots, U_T$ consumed by the inverse cumulative distribution function (inverse-CDF) method in ascending-token-ID order at every step, so that the same random draw is applied to two possibly different distributions. We record the first-disagreement step $t_{\mathrm{dis}} = \min\{t : \mathrm{token}_A(t) \neq \mathrm{token}_B(t)\}$ as a zero-indexed position, so that $t_{\mathrm{dis}} = 0$ means the two branches disagree on the first decoded token. For GSM8K we also record whether the extracted final numeric answer differs between the two branches.

4 METHOD

4.1 COMMUTATION THEORY

Lemma 1 (Renormalization commutes with power-scaling). *Let $p \in \Delta^{n-1}$ and let $S \subseteq \{1, \ldots, n\}$ satisfy $\sum_{i \in S} p_i > 0$. Define p^S as p restricted to S and renormalized, and let Q_τ^S denote the temperature operator applied within S. Then*

$$Q_\tau(p^S) = (Q_\tau(p))^S. \tag{6}$$

Temperature scaling and renormalization are therefore interchangeable whenever the support is held fixed.

Proposition 1 (Nucleus movement and top-p order dependence). *Let p be sorted as above and let $m(p)$ denote its minimal admissible nucleus index at level p_0. Temperature scaling moves that index in a direction fixed by τ alone,*

$$m(Q_\tau(p)) \leq m(p) \text{ for } \tau < 1, \quad m(Q_\tau(p)) = m(p) \text{ at } \tau = 1, \quad m(Q_\tau(p)) \geq m(p) \text{ for } \tau > 1, \tag{7}$$

for every p. The two orders agree, $N_{p_0}(Q_\tau(p)) = Q_\tau(N_{p_0}(p))$, if and only if the two indices coincide.

Proposition 2 (Bound on the top-p order effect at $\tau \leq 1$). *For any finite $p \in \Delta^{n-1}$, any $p_0 \in (0, 1]$, and any $\tau \leq 1$,*

$$\mathrm{TV}\big(N_{p_0}(Q_\tau(p)),\, Q_\tau(N_{p_0}(p))\big) \leq 1 - p_0. \tag{8}$$

The bound is specific to $\tau \leq 1$. For $\tau > 1$ the shorter of the two retained prefixes carries no guaranteed mass, and Section 5.1 shows that no bound of this form holds there.

Proposition 3 (Min-p reparameterization). *For any $\tau > 0$, $\alpha \in (0, 1]$, and any finite $p \in \Delta^{n-1}$,*

$$M_\alpha(Q_\tau(p)) = Q_\tau\big(M_{\alpha^\tau}(p)\big). \tag{9}$$

Proposition 4 (Impossibility of a top-p threshold reparameterization). *Fix $\tau > 0$ with $\tau \neq 1$, and fix $p_0 \in (0, 1)$. There is no function $f(\tau, p_0)$, independent of p, such that $N_{p_0}(Q_\tau(p)) = Q_\tau\big(N_{f(\tau,p_0)}(p)\big)$ holds for every finite categorical distribution p.*

Lemma 1 and Propositions 1–4 are proved in Appendix A. Min-p order dependence is removed by an exact reparameterization of the threshold, while top-p order dependence cannot be removed at the level of the threshold parameter alone, because for any fixed $\tau \neq 1$ and $p_0 < 1$ two distributions can be constructed whose required correction values lie in disjoint intervals, so no single $f(\tau, p_0)$ can serve both.

The two remaining rules in common use need no correction. Top-k keeps the k largest coordinates and temperature preserves the sorted order, so both compositions retain the same set and Lemma 1 makes them identical; min-k acts on the sorted logit spectrum and is invariant to temperature by construction (Ding et al., 2026). Order dependence arises for exactly the two rules that select their support from probability values, and of those only min-p admits a threshold that absorbs the temperature.

4.2 Measurement Pipeline

We evaluate these predictions on three models, LFM2.5-1.2B-Instruct, MiniCPM5-1B, and Granite-4.1-3B, which we refer to as LFM2.5, MiniCPM5, and Granite-4.1. The quantities in Section 4.1 depend on the shape of the next-token distribution and not on model scale, and these three checkpoints span a wide range of that shape: their mean top-1 probability runs from 0.434 to 0.610 and their mean cumulative mass at rank 64 from 0.870 to 0.962 (Table 5). Each model runs in bfloat16, and every operator computation is carried out in float32 over the full vocabulary, so the measured divergence is not limited by the precision of the operators themselves. Each model processes the complete WikiText-103 test split, split into non-overlapping windows of 1024 tokens with no cross-document concatenation, and a single teacher-forced forward pass over each window yields the full logit vector at every valid position. Appendix E gives the checkpoint identifiers and the tokenization and batching details. Teacher forcing on a fixed corpus holds the conditioning context identical across both orders at every position, which isolates the per-position effect of the operator composition.

At every valid position we compute the base distribution and, over the grid $\tau \in \{0.5, 0.7, 1.0, 1.5\}$, $p_0 \in \{0.9, 0.95\}$, $\alpha \in \{0.05, 0.1\}$, we evaluate both orders of the top-p pair, both orders of the uncorrected min-p pair (same α on both sides), and the corrected min-p pair (threshold α^τ applied before temperature, matched against α applied after temperature). Total variation, nucleus-index difference, retained-mass difference, support-size ratio, and the direction of nucleus movement relative to Proposition 1 are recorded at every position, together with the entropy of the base distribution. The min-p threshold is set by the ratio to the most likely token, so the size of the retained set follows the flatness of the distribution and not any fixed rank, which is why every quantity is computed over the full vocabulary.

Unless otherwise noted, detailed analyses use $\tau = 0.7$ and $p_0 = 0.95$. To separate numerical precision from the effect of operator order, we also repeat the measurement at that setting for LFM2.5 and MiniCPM5, on all evaluated positions, under three conditions relative to a higher-precision reference; Appendix B defines them and reports the result. Bootstrap confidence intervals for the distributional experiments resample whole documents with replacement, which accounts for correlation between positions inside a window and between windows drawn from the same article.

4.3 Sequence-Level Coupling

We evaluate propagation to benchmark outcomes using LFM2.5 on the full official GSM8K test set (Cobbe et al., 2021) (1319 problems) and on all 164 HumanEval problems (Chen et al., 2021). We score exact-match accuracy on the final numeric answer for GSM8K, and compute the first-disagreement step and the number of distinct completions per problem on both datasets. Token-level divergence need not produce a different benchmark score, since many different token sequences map to the same scored answer. We therefore evaluate the distributional divergence and the benchmark outcome separately. Each prompt is rendered with the model's chat template, and the exact instruction text for both datasets is reproduced in Appendix E. GSM8K prompts ask the model to end its reply with the final numeric answer in a fixed format, and answers are read from that format and normalized as described in Appendix E. Generations are capped at 512 new tokens for GSM8K and 256 for HumanEval. At $\tau = 0.7$, $p_0 = 0.95$, $\alpha = 0.1$, we generate four samples per GSM8K item and sixteen per HumanEval item, using the shared-uniform inverse-CDF coupling described above, so that the two orders are compared against the same stream of randomness at every step until their token histories diverge.

Consuming one stream of uniform variates in a fixed token order makes the two branches select the same token whenever the two distributions coincide, so the first disagreement arises from the composition order and not from independent random draws. We repeat this for three conditions, the top-p order swap, the uncorrected min-p pairing, and the α^τ-corrected min-p pairing, each run with three seeds. Paired accuracy differences are summarized with hierarchical bootstrap confidence intervals that resample both seeds and problems.

Table 1: Mean per-position total variation distance between the two composition orders over the WikiText-103 test split. The top-p columns compare temperature-then-top-p with top-p-then-temperature at each threshold p_0; the min-p columns compare the two orders at each threshold α, first with α unchanged and then with the α^τ correction, whose value is the same at both thresholds.

		Top-p		Min-p, uncorrected		Min-p
Model	τ	$p_0 = 0.9$	$p_0 = 0.95$	$\alpha = 0.05$	$\alpha = 0.1$	corrected
LFM2.5	0.5	0.0441	0.0246	0.0339	0.0456	0
LFM2.5	0.7	0.0557	0.0324	0.0385	0.0493	0
LFM2.5	1.0	0	0	0	0	0
LFM2.5	1.5	0.4484	0.3797	0.1909	0.1787	0
MiniCPM5	0.5	0.0518	0.0285	0.0452	0.0603	0
MiniCPM5	0.7	0.0635	0.0357	0.0511	0.0629	0
MiniCPM5	1.0	0	0	0	0	0
MiniCPM5	1.5	0.4961	0.4190	0.2306	0.2135	0
Granite-4.1	0.5	0.0353	0.0197	0.0292	0.0367	0
Granite-4.1	0.7	0.0413	0.0244	0.0315	0.0433	0
Granite-4.1	1.0	0	0	0	0	0
Granite-4.1	1.5	0.2708	0.2308	0.1482	0.1370	0

5 Results

5.1 Non-Commutation of Temperature and Top-p

The top-p columns of Table 1 give the mean total variation distance between the two orders for each of the three models across the full (τ, p_0) grid. Across roughly 290,000 evaluated positions per model, the mean distance is exactly 0 at $\tau = 1$ for all three models, where Q_1 is the identity and the two orders coincide. Total variation is positive at every other tested τ, but it does not increase monotonically with $|\tau - 1|$. For $\tau < 1$, the value at $\tau = 0.7$ is larger than at $\tau = 0.5$ (for example, 0.0557 against 0.0441 at $p_0 = 0.9$ for LFM2.5, and the same ordering holds for the other two models).

Proposition 2 predicts this shape. Its proof shows that the distance equals $1 - S/T$, where S is the mass the tempered nucleus retains under Q_τ. As τ falls toward zero the tempered distribution concentrates on its largest coordinates, so S approaches 1 and the distance returns to zero. It is zero at $\tau = 1$ as well, so on $(0, 1]$ it must rise and fall in between, and the measurements place the peak at $\tau = 0.7$ for every model and threshold.

The largest divergence among all tested settings occurs at $\tau = 1.5$ in every case, reaching 0.4484, 0.4961, and 0.2708 at $p_0 = 0.9$ for LFM2.5, MiniCPM5, and Granite-4.1 respectively, with a 95% document-bootstrap interval of $[0.4447, 0.4520]$ for LFM2.5. These three values exceed $1 - p_0$, so the bound of Proposition 2 does not extend above unit temperature. The direction in which the nucleus moves agrees with Proposition 1 at every position, in each of the 24 tested settings. Table 6 in Appendix F lists every setting with its interval, its direction-agreement rate, and the accompanying support statistics.

5.2 Order Independence of the Corrected Min-p Pairing

The uncorrected pairing holds α fixed across both orders, and its total variation is nonzero at every grid point away from $\tau = 1$ for every model, reaching 0.2306 for MiniCPM5 at $\tau = 1.5$, $\alpha = 0.05$ (Table 1). As with the top-p pairing, this divergence is not a strictly monotone function of $|\tau - 1|$; the value at $\tau = 0.7$ exceeds the value at $\tau = 0.5$ for several model and α combinations, while $\tau = 1.5$ remains the largest in every case. The α^τ-corrected pairing gives a total variation of zero, to eight decimal places, at every one of the 24 tested (model, τ, α) combinations, and the two orders retain identical supports at every evaluated position. This includes settings where the uncorrected pairing shows a divergence larger than 0.2, and it holds across roughly $870{,}000$ next-token distributions.

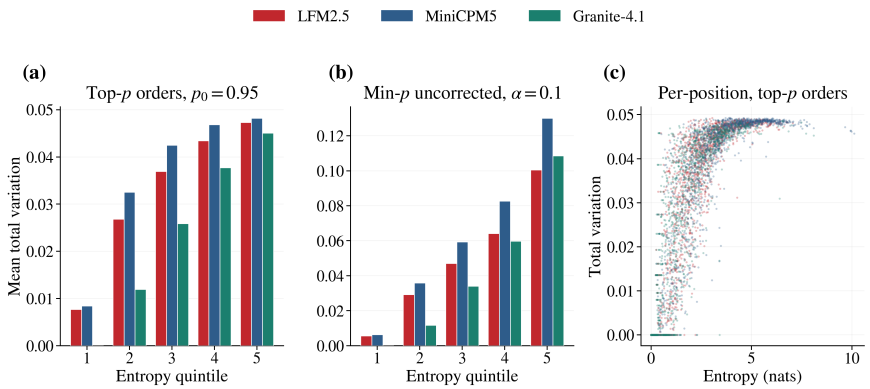

Figure 1: Total variation distance at $\tau = 0.7$ stratified by entropy quintile of the base distribution, for the three models. Mean per-position distance in each quintile for (a) the two top-p orders at $p_0 = 0.95$ and (b) the uncorrected min-p pairing at $\alpha = 0.1$. (c) Per-position scatter of base-distribution entropy in nats against the top-p total variation distance, with one color per model.

The two sides follow different computational paths, one selecting its support before temperature and the other after, with different intermediate values and a different renormalization. Their agreement therefore shows that the identity survives finite-precision arithmetic and does not reflect a shared computation.

A precision and vocabulary-restriction ablation, reported in Appendix B, compares these divergences with the numerical error incurred in measuring them. Reduced precision and a top-1024 restriction each shift the measured total variation by less than 0.002, more than an order of magnitude below the mean divergence at the same setting, 0.0324 for LFM2.5 and 0.0357 for MiniCPM5. The top-1024 restriction also leaves the nucleus incomplete at 10.55% of positions for LFM2.5 and 26.57% for MiniCPM5, so a rank-limited measurement would compare incomplete supports.

5.3 Entropy Concentration of the Divergence

The divergence concentrates at high-entropy positions (Figure 1). In the lowest entropy quintile, where next-token distributions are most concentrated, the total variation between the two top-p orders is small, for example 0.0077 for LFM2.5, and is exactly 0 for Granite-4.1, whose mean entropy in that quintile is only 0.043 nats. Divergence increases across quintiles for both the top-p pair and the uncorrected min-p pair, reaching 0.0473 and 0.1006 respectively in the highest entropy quintile for LFM2.5, while the corrected min-p pairing remains exactly zero in every quintile for every model. Figure 1(c) shows the same pattern at the level of individual positions, with near-zero-entropy positions clustered at zero divergence for every model.

Every measured position stays below $1 - p_0$, the bound of Proposition 2, which at $p_0 = 0.95$ places the whole scatter at or below 0.05. The largest per-position value is 0.0493, within 1.4% of the bound, so Proposition 2 is close to tight on natural next-token distributions. Entropy is a property of the base distribution alone, so this profile can be measured for a model and a task without running either order, and it indicates in advance where the choice of order can take effect.

5.4 Default Order in the Transformers Library

With temperature, top-p, and min-p all enabled, the Transformers library (Wolf et al., 2020) (version 5.9.0) orders the processors as temperature, then top-p, then min-p for each of the three checkpoints. That order is fixed inside the library and is not among the sampling parameters a caller sets. The same processor classes can also be composed in the opposite sequence, and on a constructed input at $\tau = 0.5$, $p_0 = 0.9$ the two compositions give a total variation distance of 0.0303. Appendix C gives the procedure and the full result.

Table 2: Sequence-level coupled propagation on GSM8K. Columns give the full-agreement rate of the coupled completion pairs, the exact-match accuracy of each order, temperature first and truncation first, as mean $\pm$ standard deviation across three seeds, and the paired accuracy difference with its 95% hierarchical bootstrap confidence interval over seeds and problems.

Condition	Agreement	Temp.-first acc.	Trunc.-first acc.	Difference [95% CI]
Top-p order swap	32.6%	0.4712 ± 0.0017	0.4578 ± 0.0026	$+0.0134$ $[0.0035, 0.0231]$
Min-p uncorrected	22.3%	0.4689 ± 0.0016	0.4670 ± 0.0014	$+0.0019$ $[-0.0075, 0.0116]$
Min-p corrected	100%	0.4689 ± 0.0016	0.4689 ± 0.0016	0 $[0, 0]$

5.5 Propagation to a Sequence-Level Benchmark

Figure 2 in Appendix F shows the distribution of the first-disagreement step under the shared-uniform coupling. On GSM8K, the top-p order swap produces a full-agreement rate of 32.6% over coupled sample pairs pooled across problems and seeds, with a median first-disagreement step of 115 tokens; the uncorrected min-p pairing gives a full-agreement rate of 22.3% and a median first-disagreement step of 92 tokens. The α^τ-corrected min-p pairing shows full agreement in every coupled pair.

A single mismatch at any step would separate the two branches for the rest of the generation, so full agreement requires the identity of Proposition 3 to hold at every one of up to 512 sequential applications of the operator pair, not only at independent positions. The agreement rate and the first-disagreement step describe this coupling, since a different coupling of the same two distributions can place the first disagreement elsewhere. The accuracy comparison below does not depend on that choice, because each order is scored against its own marginal distribution and the coupling only removes the variance from independent draws.

The top-p order swap also changes GSM8K exact-match accuracy (Table 2). Mean accuracy across three seeds is 0.4712 for temperature-then-top-p and 0.4578 for top-p-then-temperature. The paired difference is $+0.0134$, with a 95% hierarchical bootstrap confidence interval (CI) of $[0.0035, 0.0231]$ over seeds and problems that excludes zero, The two arms differ only in the composition order, so at this setting that choice accounts for a change of 1.34 percentage points in exact-match accuracy. Proposition 1 fixes which order truncates more tightly here: at $\tau = 0.7$ the tempered nucleus is the shorter prefix, so temperature-then-top-p applies the stronger truncation, and that is the order with the higher score. The uncorrected min-p pairing gives a paired difference whose interval includes zero, despite disagreeing on more completions than the top-p swap. The corrected min-p pairing gives a paired difference of 0 with a confidence interval of $[0, 0]$, since the two orders induce identical distributions and therefore identical completions under the coupling.

Beyond the scored answer, the two orders also produce different numbers of distinct completions per problem on HumanEval, 11.33 against 11.79 for the top-p order swap and 10.63 against 11.46 for the uncorrected min-p pairing, while the corrected pairing gives 10.63 for both. Appendix D reports these counts in full. These differences change the sample set on which self-consistency (Wang et al., 2023) and pass@k (Chen et al., 2021) are computed, so composition order can affect such evaluations without changing any individual scored answer.

6 Discussion

The two truncation rules behave differently because their thresholds depend on different features of the distribution. The min-p threshold depends only on the ratio of a candidate token's probability to the maximum probability, and raising every probability to the power $1/\tau$ raises that ratio to the same power, so a single change of the threshold compensates for the change exactly. The top-p threshold depends on a cumulative sum over the whole sorted distribution, and temperature changes that cumulative mass in a distribution-dependent way, so no function of τ and p_0 alone can follow it. Proposition 2 accounts for the asymmetry between $\tau < 1$ and $\tau > 1$ in Table 1. Below unit temperature the tempered nucleus is the shorter of the two prefixes and still carries at least p_0 of the tempered mass, which caps the disagreement at $1 - p_0$; above unit temperature the roles reverse,

the shorter prefix carries no guaranteed mass, and the measured divergence rises by an order of magnitude.

A tuple such as "temperature 0.7, top-p 0.95" names two operations but not their order, and with top-p the two orders can sample from different distributions once $\tau \neq 1$. No adjustment of p_0 that depends only on τ and p_0 removes the difference for every distribution, so the order is an additional configuration choice and not a consequence of the two numbers already reported. The min-p case admits a closed-form correction. Applying the threshold α^τ before temperature scaling gives the same distribution as applying α after it, so τ, α, and the order together determine the distribution, and a system that adopts the corrected threshold makes the order immaterial. The correction changes one scalar and leaves the sampling procedure untouched, and because Proposition 3 holds for every finite distribution it can be checked against an implementation on constructed inputs, without running a model. When the order does matter, its effect is concentrated at decoding positions with several plausible continuations. It is not confined to extreme temperatures: at $\tau = 0.7$, a setting in common use, the two top-p orders differ by 0.0557 on average for LFM2.5 at $p_0 = 0.9$. The sequence-level experiment runs at $\tau = 0.7$ and $p_0 = 0.95$, where Proposition 2 caps the per-step distance at 0.05 and the measured mean is 0.0324. A per-step difference of that size still moves exact-match accuracy by 1.34 percentage points over a 512-token horizon.

The corrected min-p pairing gives matching distributions, identical coupled completions, and identical GSM8K accuracy. Where the effect remains, the size of the per-step divergence and the size of the change it produces in a benchmark score need not track each other. The top-p order swap shifts GSM8K accuracy by more than one percentage point, while the uncorrected min-p pairing changes most completions and leaves an accuracy difference whose interval includes zero, since two branches can diverge in wording and still reach the same final numeric answer. The entropy stratification in Figure 1 is consistent with this contrast.

The propositions hold for every finite categorical distribution, so what varies across models is the size of the effect and not whether it arises. The measurements fix that size on the distributions three current models produce. The library study documents a specific default processor order in Transformers $5.9.0$, and Propositions 1 and 4 apply to any implementation that composes these operators in sequence. Taken with the distribution-level measurements, the GSM8K result shows that composition order is an implementation-level choice with measurable consequences for reproducible stochastic evaluation.

7 Conclusion

A reported sampling configuration such as temperature 0.7 with top-p 0.95 does not, by itself, determine the distribution a language model samples from, because the order in which temperature and truncation are applied is left unspecified. For top-p, the two orders can differ once $\tau \neq 1$, and no adjustment of the threshold alone removes the difference for every distribution, while for min-p the two orders coincide once α is replaced by α^τ. The predicted identity and the order effects are verified numerically on three language models over the complete vocabulary, the same processor classes of the Transformers library can be composed in either order, and on GSM8K the choice of top-p order shifts exact-match accuracy by a statistically significant 1.34 percentage points. The composition order therefore belongs in a reported configuration alongside τ and p_0, together with the inference implementation and its version, since that is what fixes the order; where min-p is used, the corrected threshold removes the ambiguity outright.

AI use statement

Generative AI tools were used to assist with editing prose, with writing the experiment and plotting code, and with formatting LaTeX tables, figures, and captions from the underlying numerical results. They were not used to generate the propositions, the proofs, the experimental design, or the analysis of results. All AI-assisted text was reviewed and edited by the authors, who take responsibility for the final content of this work.

Reproducibility statement

The formal statements in Section 4.1 are proved in full in Appendix A. The measurement pipeline, model checkpoints, dataset splits, hyperparameter grid, prompt templates, and coupling procedure are specified in Section 4 and Appendix E. All datasets used (WikiText-103, GSM8K, HumanEval) and all three model checkpoints are publicly available, and the Transformers library version used for the library study (5.9.0) is recorded in Appendix E.

References

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 *Advances in Neural Information Processing Systems*, volume 38, pp. 32287–32318, 2025. doi: 10.52202/085713-0958.

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.

Luka Borec, Philipp Sadler, and David Schlangen. The unreasonable ineffectiveness of nucleus sampling on mitigating text memorization. In *Proceedings of the 17th International Natural Language Generation Conference*, pp. 358–370. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.inlg-main.30.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.681.

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. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1082.

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

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9992–10020. Association for Computational Linguistics, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

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. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.249.

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.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3752–3762. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1365.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023. doi: 10.1162/tacl_a_00536.

Moin Nadeem, Tianxing He, Kyunghyun Cho, and James Glass. A systematic characterization of sampling algorithms for open-ended language generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 334–346. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.aacl-main.36.

Nhat Minh Nguyen, 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 *International Conference on Learning Representations*, 2025.

Rylan Schaeffer, Joshua Kazdan, and Yegor Denisov-Blanch. Min-p, max exaggeration: A critical analysis of Min-p sampling in language models. *arXiv preprint arXiv:2506.13681*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022. doi: 10.1162/tacl_a_00502.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

A PROOFS

A.1 PROOF OF LEMMA 1

Let $Z = \sum_{i \in S} p_i$. By definition, $p_i^S = p_i/Z$ for $i \in S$ and 0 otherwise. Applying Q_τ within S,

$$Q_\tau(p^S)_i = \frac{(p_i/Z)^{1/\tau}}{\sum_{j \in S}(p_j/Z)^{1/\tau}} = \frac{p_i^{1/\tau} Z^{-1/\tau}}{Z^{-1/\tau} \sum_{j \in S} p_j^{1/\tau}} = \frac{p_i^{1/\tau}}{\sum_{j \in S} p_j^{1/\tau}},$$

for $i \in S$, and 0 otherwise. This expression does not depend on Z, so it equals $(Q_\tau(p))_i^S$, which is $Q_\tau(p)_i$ restricted to S and renormalized by $\sum_{j \in S} Q_\tau(p)_j = \sum_{j \in S} p_j^{1/\tau} / \sum_{k=1}^n p_k^{1/\tau}$; the global normalizer $\sum_k p_k^{1/\tau}$ cancels in the renormalization, giving the same expression. Hence $Q_\tau(p^S) = (Q_\tau(p))^S$.

A.2 PROOF OF PROPOSITION 1

For $\tau < 1$, $1/\tau > 1$, so $g(x) = x^{1/\tau - 1}$ is a strictly increasing function of $x \geq 0$. Restrict attention to indices with $p_i > 0$; any index with $p_i = 0$ satisfies $Q_\tau(p)_i = 0^{1/\tau} = 0$ by convention as well, so it contributes zero to every prefix sum on both sides and can be dropped from the argument without loss of generality. Write $r_i = Q_\tau(p)_i / p_i = p_i^{1/\tau - 1} / \sum_j p_j^{1/\tau}$. Since p is sorted non-increasingly, r is non-increasing, and $\sum_i p_i r_i = 1$. Fix a prefix length m and set $A = \sum_{i \leq m} p_i$. Using $r_i \geq r_m$ for $i \leq m$ and $r_i \leq r_m$ for $i > m$,

$$\begin{aligned}\sum_{i \leq m} p_i r_i - A &= (1 - A) \sum_{i \leq m} p_i r_i - A \sum_{i > m} p_i r_i \\ &\geq r_m \Big[(1 - A) \sum_{i \leq m} p_i - A \sum_{i > m} p_i \Big] = r_m \big[(1 - A)A - A(1 - A) \big] = 0,\end{aligned}$$

so every prefix of $Q_\tau(p)$ carries at least as much mass as the same prefix of p. Hence if $\sum_{i \leq m} p_i \geq p_0$ for some m, the same holds under $Q_\tau(p)$, and the minimal index achieving the threshold satisfies $m(Q_\tau(p)) \leq m(p)$. For $\tau > 1$, $1/\tau - 1 < 0$ makes r non-decreasing, and the same computation with the inequalities reversed gives $m(Q_\tau(p)) \geq m(p)$; at $\tau = 1$, Q_1 is the identity, so $m(Q_1(p)) = m(p)$. Cumulative mass exactly equal to p_0 is resolved by the inequality $\geq$ in the definitions of $m(\cdot)$ and $S_\alpha(\cdot)$, which select the same index set on either side. Coordinates of equal probability are ordered by the ascending-token-ID rule, and temperature scaling preserves both that order and the equalities, since $x \mapsto x^{1/\tau}$ is strictly increasing.

Finally, $N_{p_0}(Q_\tau(p))$ and $Q_\tau(N_{p_0}(p))$ both truncate p to a prefix and then apply temperature and renormalization within that prefix (using Lemma 1 for the second expression, since $N_{p_0}(p)$ is exactly p restricted to its own nucleus and renormalized, and Q_τ then acts within that fixed support). The two expressions therefore agree when $m(Q_\tau(p)) = m(p)$. Conversely, suppose $m_Q := m(Q_\tau(p)) \neq m := m(p)$, and take $m_Q < m$ without loss of generality. Minimality of m gives $\sum_{i < m} p_i < p_0 \leq \sum_{i \leq m} p_i$, so $p_m > 0$ and hence $Q_\tau(p)_m > 0$. The second expression assigns positive mass to index m while the first assigns zero, so the two differ. The equality holds if and only if the two minimal nucleus indices coincide.

A.3 PROOF OF PROPOSITION 2

Let $\tau \leq 1$, so $m_Q := m(Q_\tau(p)) \leq m := m(p)$ by Proposition 1. By Lemma 1, $Q_\tau(N_{p_0}(p))$ is $Q_\tau(p)$ restricted to the first m indices and renormalized, and $N_{p_0}(Q_\tau(p))$ is $Q_\tau(p)$ restricted to the first m_Q indices and renormalized. Write $S = \sum_{i \leq m_Q} Q_\tau(p)_i$ and $T = \sum_{i \leq m} Q_\tau(p)_i$, so that $0 < S \leq T \leq 1$. Then

$$\mathrm{TV} = \frac{1}{2} \Big[\sum_{i \leq m_Q} Q_\tau(p)_i \Big(\frac{1}{S} - \frac{1}{T} \Big) + \sum_{m_Q < i \leq m} \frac{Q_\tau(p)_i}{T} \Big] = \frac{1}{2} \Big[\Big(1 - \frac{S}{T} \Big) + \frac{T - S}{T} \Big] = 1 - \frac{S}{T}.$$

By the definition of m_Q as the minimal admissible nucleus index of $Q_\tau(p)$ at level p_0, we have $S \geq p_0$, and $T \leq 1$ gives $S/T \geq S$. Hence $\mathrm{TV} \leq 1 - S \leq 1 - p_0$. For $\tau > 1$ the two supports

exchange roles, the smaller one is the prefix of length m, and its mass under $Q_\tau(p)$ has no lower bound in terms of p_0, so the argument gives no bound.

A.4 Proof of Proposition 3

By definition, $i \in S_\alpha(Q_\tau(p))$ if and only if $Q_\tau(p)_i \geq \alpha \max_j Q_\tau(p)_j$. Since $Q_\tau(p)_i = p_i^{1/\tau} / \sum_k p_k^{1/\tau}$, the normalizing constant cancels in this inequality, leaving $p_i^{1/\tau} \geq \alpha \max_j p_j^{1/\tau} = \alpha(\max_j p_j)^{1/\tau}$ (using that $x \mapsto x^{1/\tau}$ is increasing, so the maximizer is unchanged). Raising both sides to the power $\tau > 0$, which preserves the inequality since both sides are non-negative, gives $p_i \geq \alpha^\tau \max_j p_j$, which is exactly the condition $i \in S_{\alpha^\tau}(p)$. Hence $S_\alpha(Q_\tau(p)) = S_{\alpha^\tau}(p)$ as index sets. Both $M_\alpha(Q_\tau(p))$ and $Q_\tau(M_{\alpha^\tau}(p))$ restrict p to this common support and apply Q_τ within it (using Lemma 1 again for the right-hand side), so the two expressions are identical.

A.5 Proof of Proposition 4

Specific counterexample. We first give a concrete counterexample at $\tau = 0.5$ and $p_0 = 0.7$, then generalize it to arbitrary $\tau \neq 1$ and $p_0 \in (0, 1)$.

Suppose, for contradiction, that some value $f = f(0.5, 0.7)$, independent of p, satisfies $N_{0.7}(Q_{0.5}(p)) = Q_{0.5}(N_f(p))$ for every finite three-token distribution p. Since $\tau = 0.5$ gives $1/\tau = 2$, $Q_{0.5}$ squares each coordinate and renormalizes.

Consider $p = (0.53, 0.235, 0.235)$, whose cumulative mass at rank 1 is $C_1(p) = 0.53$. Squaring the coordinates and normalizing gives $Q_{0.5}(p) \approx (0.71779, 0.14110, 0.14110)$. The cumulative mass at rank 1 under $Q_{0.5}(p)$ is $0.71779 \geq 0.7$, so $m(Q_{0.5}(p)) = 1$. For the equality $N_{0.7}(Q_{0.5}(p)) = Q_{0.5}(N_f(p))$ to hold at this p, $N_f(p)$ must select the same prefix index, 1, when applied to the untempered p. Since $C_1(p) = 0.53$, this requires

$$f \in (0,\ 0.53].$$

Now consider $p' = (0.58, 0.40, 0.02)$, whose cumulative masses are $C_1(p') = 0.58$ and $C_2(p') = 0.98$. Squaring and normalizing gives $Q_{0.5}(p') \approx (0.67713, 0.32206, 0.00081)$. The cumulative mass at rank 1 under $Q_{0.5}(p')$ is $0.67713 < 0.7$, and at rank 2 it is $0.99919 \geq 0.7$, so $m(Q_{0.5}(p')) = 2$. For the equality to hold at p', $N_f(p')$ must select prefix index 2 when applied to the untempered p', requiring

$$f \in (0.58,\ 0.98].$$

The two required intervals, $(0, 0.53]$ and $(0.58, 0.98]$, are disjoint. No single value f can lie in both, which contradicts the assumption that f depends only on (τ, p_0), establishing the claim for this specific pair.

General construction for arbitrary $\tau \neq 1$ and $p_0 \in (0, 1)$. Fix $\tau \neq 1$ and $p_0 \in (0, 1)$. Two families of distributions require disjoint threshold intervals.

The first family is uniform. For a uniform distribution $u^{(n)}$ on n coordinates, $Q_\tau(u^{(n)}) = u^{(n)}$ for every τ, so both orders truncate at the same rank $m_n = \lceil p_0 n \rceil$, and the equality $N_{p_0}(Q_\tau(u^{(n)})) = Q_\tau(N_f(u^{(n)}))$ holds exactly when N_f selects that same rank, which requires

$$f \in \left(\tfrac{m_n - 1}{n},\ \tfrac{m_n}{n}\right], \qquad \tfrac{m_n - 1}{n} \geq p_0 - \tfrac{1}{n}, \qquad \tfrac{m_n}{n} \leq p_0 + \tfrac{1}{n}.$$

The second family consists of distributions p with one head coordinate of mass c and k tail coordinates of mass $(1 - c)/k$ each, with $k > (1 - c)/c$ so that the head is the largest coordinate. The tempered mass at rank 1 is

$$g_\tau(c, k) = \frac{c^{1/\tau}}{c^{1/\tau} + k^{\,1-1/\tau}(1 - c)^{1/\tau}}.$$

For $\tau < 1$ the exponent $1 - 1/\tau$ is negative, so $g_\tau(c, k) \to 1$ as $k \to \infty$ and $m(Q_\tau(p)) = 1$ for all large k, which requires $f \in (0, c]$. For $\tau > 1$ the exponent is positive, so $g_\tau(c, k) \to 0$ as $k \to \infty$

and $m(Q_\tau(p)) \geq 2$ for all large k, which requires $f \in (C_{m-1}(p), C_m(p)]$ with $m = m(Q_\tau(p))$ and therefore $f > C_1(p) = c$.

The two intervals are now separated by a choice of n and c. For $\tau < 1$, take $n > 2/p_0$ and $c = p_0/2$. The uniform family requires $f \geq p_0 - 1/n > p_0/2 = c$, while the head-and-tail family requires $f \leq c$, so the two intervals are disjoint. For $\tau > 1$, take $n > 2/(1 - p_0)$ and $c = p_0 + (1 - p_0)/2$. The uniform family requires $f \leq p_0 + 1/n < c$, while the head-and-tail family requires $f > c$, and the intervals are again disjoint. In both cases no single value of f serves both distributions, so no function $f(\tau, p_0)$ independent of p can satisfy the required equality for every p. The three-token example above is the case $\tau = 0.5$, $p_0 = 0.7$, where the same separation is already visible at $n = 3$.

The excluded boundary $p_0 = 1$ is different. At $p_0 = 1$, $N_1(p) = p$ for every p, so both $N_1(Q_\tau(p))$ and $Q_\tau(N_1(p))$ equal $Q_\tau(p)$, and $f(\tau, 1) = 1$ satisfies the required equality.

The result rules out reparameterizing the top-p threshold as $p_0 \mapsto f(\tau, p_0)$ at fixed τ and fixed $p_0 < 1$.

B Precision and Vocabulary-Restriction Ablation

Table 3 reports the precision and vocabulary-restriction ablation, obtained at $\tau = 0.7$, $p_0 = 0.95$ against a float32-weight, float32-forward, full-vocabulary reference, for LFM2.5 and MiniCPM5. Running the entire forward pass with bfloat16 weights introduces a mean absolute total variation error of 0.00186 for LFM2.5 and 0.00126 for MiniCPM5 relative to the float32 reference. In both models, this exceeds the error from only rounding the final float32 logits to bfloat16 precision after an otherwise float32 forward pass (0.00120 and 0.00086 respectively), so bfloat16 execution adds error beyond final-logit rounding alone. Restricting the measurement to the top 1024 logits gives a mean absolute error of 0.00097 for LFM2.5 and 0.00171 for MiniCPM5. The top-1024 restriction also produces an incomplete nucleus for 10.55% of positions for LFM2.5 and 26.57% for MiniCPM5, so at those positions the true top-p nucleus extends beyond rank 1024 and a measurement restricted to that many logits compares supports that are themselves incomplete. All three error sources remain more than an order of magnitude below the quantity they are measured against, since the reference mean total variation at this setting is 0.0324 for LFM2.5 and 0.0357 for MiniCPM5 (Table 3).

Table 3: End-to-end precision and vocabulary-restriction error at $\tau = 0.7$, $p_0 = 0.95$, relative to a float32-weight, float32-forward, full-vocabulary reference. Columns give the number of evaluated positions, the reference mean total variation, the mean absolute error under a bfloat16 forward pass, under bfloat16 rounding of the final logits only, and under a float32 top-1024 restriction, and the fraction of positions whose nucleus that restriction leaves incomplete.

	LFM2.5	MiniCPM5
Evaluated positions	296,205	285,180
Reference mean TV	0.03244	0.03570
Error, bfloat16 weights	0.001858	0.001263
Error, bfloat16 logits only	0.001200	0.000864
Error, float32 top-1024	0.000967	0.001705
Incomplete top-1024 nucleus	10.55%	26.57%

C Library Execution Study

We evaluate both orders with the temperature, top-p, and min-p logits processors of the Transformers library (Wolf et al., 2020) (version 5.9.0), listed by class name in Appendix E. We inspect the processor sequence that the library instantiates for each checkpoint, then compose the same processors in the opposite sequence. We compare the two resulting distributions on an illustrative six-token example.

The library constructs a single order for all three checkpoints by default, temperature followed by top-p followed by min-p. On the six-token example at $\tau = 0.5$, $p_0 = 0.9$, the two compositions retain supports of size 4 and 5 and give a total variation distance of 0.0303, so the same processor classes support either order.

D Sample Diversity Under the Two Orders

Table 4: Mean number of distinct completions per problem under each composition order, across three seeds. GSM8K draws four samples per problem and HumanEval sixteen.

	GSM8K		HumanEval	
Condition	Temperature first	Truncation first	Temperature first	Truncation first
Top-p order swap	4.00	4.00	11.33	11.79
Min-p, uncorrected	4.00	4.00	10.63	11.46
Min-p, corrected	4.00	4.00	10.63	10.63

The two orders produce different numbers of distinct completions per problem (Table 4). On GSM8K all four samples per problem are distinct for both orders in every condition. On HumanEval, where sixteen samples are drawn per problem, the top-p order swap gives 11.33 distinct completions for temperature-then-top-p against 11.79 for top-p-then-temperature, and the uncorrected min-p pairing gives a larger gap, 10.63 against 11.46. The corrected min-p pairing gives 10.63 for both orders.

E Additional Experimental Details

The three checkpoints are LiquidAI/LFM2.5-1.2B-Instruct, openbmb/MiniCPM5-1B, and ibm-granite/granite-4.1-3b. They are loaded with bfloat16 weights, and their forward pass is computed in bfloat16 for the distributional experiments in Section 5. Logits are cast to float32 immediately after the forward pass, and all downstream operator computations (sorting, softmax, temperature scaling, cumulative summation) are carried out in float32 over the full vocabulary and accumulated in double precision. The error associated with bfloat16 model execution is measured in the precision ablation (Table 3), which uses a float32-weight, float32-forward reference. Documents are separated at article headers, tokenized independently with each model's own tokenizer, and prefixed with the model's documented beginning-of-sequence token where applicable. WikiText-103 windows are constructed with size 1024 tokens and no overlap, windows shorter than two tokens are discarded, and forward passes are batched eight windows at a time. Per-position statistics are accumulated in double precision. The distributional grid covers $\tau \in \{0.5, 0.7, 1.0, 1.5\}$, $p_0 \in \{0.9, 0.95\}$, and $\alpha \in \{0.05, 0.1\}$, applied uniformly across all three models. The split yields $296{,}205$ evaluated positions for LFM2.5, $285{,}180$ for MiniCPM5, and $289{,}229$ for Granite-4.1. The top-1024 condition of the precision ablation retains only the 1024 highest-probability tokens of the base distribution and re-derives both operator orders from that truncated set, reproducing the comparison a top-k-limited measurement would make.

The sequence-level experiments use LiquidAI/LFM2.5-1.2B-Instruct, with prompts formatted by the tokenizer's chat template (`apply_chat_template`) and a maximum of 512 new tokens for GSM8K and 256 for HumanEval. The exact instruction text appended to every GSM8K question is: "Solve this problem step by step. End your reply with the final numeric answer on its own line in exactly this format: #### <answer>". The exact instruction text used for every HumanEval prompt is: "Complete the following Python function. Reply with one Python code block containing the complete function implementation, including the signature and any needed imports.", followed by a Markdown Python code fence containing the official function stub. GSM8K answers are read from the last #### line, following the answer format of the official GSM8K evaluation script, and are then normalized by removing commas and currency symbols, rounding a value to the nearest integer when it lies within 10^{-9} of one, and otherwise formatting it to six significant digits.

Three seeds (11, 23, 37) are used for each of the six (dataset, condition) combinations. GSM8K accuracy is first averaged over the four samples of each problem, so the problem is the resampling unit. Paired hierarchical bootstrap confidence intervals then use 10,000 resamples that draw seeds and problems independently with replacement; first-disagreement statistics are computed over all generated sample pairs pooled across seeds, restricted to pairs that eventually disagree, with the full-agreement rate reported separately. The first-disagreement step is recorded as a zero-indexed token position, and agreement is evaluated over the decoded output text, which excludes terminal control tokens. Distinct-completion counts use the same decoded text.

Table 5 reports the mean cumulative base-distribution mass at fixed ranks for each model. MiniCPM5 accumulates mass more slowly at low ranks (54.7% at rank 2, against 72.8% for Granite-4.1), which corresponds to the larger support-size increase under temperature at $\tau = 1.5$ in Figure 3.

The library study (Appendix C) uses Transformers 5.9.0, and the processor sequence is the one returned by the library's generation pipeline for each checkpoint, and the corresponding classes are `TemperatureLogitsWarper`, `TopPLogitsWarper`, and `MinPLogitsWarper`.

Table 5: Mean cumulative base-distribution mass at fixed ranks over the WikiText-103 test split, per model.

Model	Rank 1	Rank 2	Rank 4	Rank 8	Rank 16	Rank 64	Rank 1024
LFM2.5	0.508	0.637	0.739	0.811	0.860	0.921	0.981
MiniCPM5	0.434	0.547	0.645	0.724	0.786	0.870	0.961
Granite-4.1	0.610	0.728	0.814	0.874	0.915	0.962	0.995

F Additional Figures and Tables

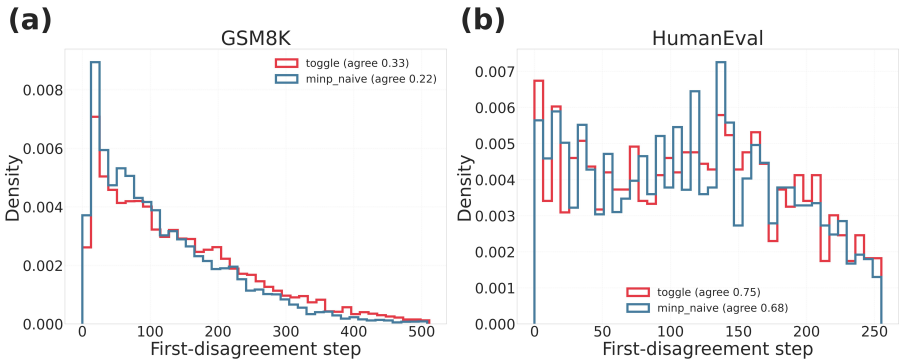

Figure 2: Density of the first-disagreement step under the shared-uniform inverse-CDF coupling, computed over the coupled generation pairs that eventually disagree, for the top-p order swap (red) and the uncorrected min-p pairing (blue) on (a) GSM8K with a 512-token horizon and (b) HumanEval with a 256-token horizon. The legend reports the full-agreement rate of each condition.

Figure 3 plots the support-size ratio and the mean nucleus-index gap against τ at $p_0 = 0.95$. The support ratio crosses 1 exactly at $\tau = 1$ and grows sharply for $\tau > 1$, consistent with a temperature above 1 flattening the distribution and admitting more tokens into the nucleus before renormalization; the growth is largest for MiniCPM5, whose base distribution has the least concentrated top-1 mass among the three models (Appendix Table 5). As with total variation, the nucleus-index gap is zero and the support-size ratio is one at $\tau = 1$, and both are largest at $\tau = 1.5$, without being a simple monotone function of $|\tau - 1|$ in between.

Table 6: Mean per-position total variation distance between the two top-p orders for all tested (model, τ, p_0) combinations, with its 95% document-bootstrap confidence interval. Rows at $\tau = 1$, where the distance is exactly zero, are omitted.

Model	τ	p_0	Mean TV	95% CI
LFM2.5	0.5	0.9	0.0441	[0.0437, 0.0445]
LFM2.5	0.5	0.95	0.0246	[0.0244, 0.0248]
LFM2.5	0.7	0.9	0.0557	[0.0553, 0.0562]
LFM2.5	0.7	0.95	0.0324	[0.0322, 0.0327]
LFM2.5	1.5	0.9	0.4484	[0.4447, 0.4520]
LFM2.5	1.5	0.95	0.3797	[0.3768, 0.3825]
MiniCPM5	0.5	0.9	0.0518	[0.0512, 0.0524]
MiniCPM5	0.5	0.95	0.0285	[0.0281, 0.0288]
MiniCPM5	0.7	0.9	0.0635	[0.0628, 0.0642]
MiniCPM5	0.7	0.95	0.0357	[0.0353, 0.0361]
MiniCPM5	1.5	0.9	0.4961	[0.4935, 0.4987]
MiniCPM5	1.5	0.95	0.4190	[0.4174, 0.4206]
Granite-4.1	0.5	0.9	0.0353	[0.0347, 0.0360]
Granite-4.1	0.5	0.95	0.0197	[0.0194, 0.0201]
Granite-4.1	0.7	0.9	0.0413	[0.0406, 0.0421]
Granite-4.1	0.7	0.95	0.0244	[0.0240, 0.0248]
Granite-4.1	1.5	0.9	0.2708	[0.2658, 0.2760]
Granite-4.1	1.5	0.95	0.2308	[0.2268, 0.2350]

Table 7: Support statistics for the two top-p orders at the same settings. Columns give the mean nucleus-index gap, the mean base-mass difference, and the mean support-size ratio. The direction-agreement rate with respect to Proposition 1 is 1 in every row.

Model	τ	p_0	Mean $\lvert m_A - m_B \rvert$	Mean mass diff.	Mean support ratio
LFM2.5	0.5	0.9	126	0.2420	0.300
LFM2.5	0.5	0.95	372	0.2385	0.210
LFM2.5	0.7	0.9	121	0.1544	0.360
LFM2.5	0.7	0.95	362	0.1390	0.260
LFM2.5	1.5	0.9	6,910	0.0828	849
LFM2.5	1.5	0.95	14,115	0.0440	1,172
MiniCPM5	0.5	0.9	427	0.2974	0.242
MiniCPM5	0.5	0.95	1,239	0.2904	0.174
MiniCPM5	0.7	0.9	410	0.1880	0.293
MiniCPM5	0.7	0.95	1,196	0.1659	0.213
MiniCPM5	1.5	0.9	23,470	0.0851	2,615
MiniCPM5	1.5	0.95	42,613	0.0445	3,488
Granite-4.1	0.5	0.9	52	0.1742	0.493
Granite-4.1	0.5	0.95	121	0.1674	0.416
Granite-4.1	0.7	0.9	48	0.1025	0.567
Granite-4.1	0.7	0.95	113	0.0888	0.491
Granite-4.1	1.5	0.9	1,210	0.0642	129
Granite-4.1	1.5	0.95	2,739	0.0352	243

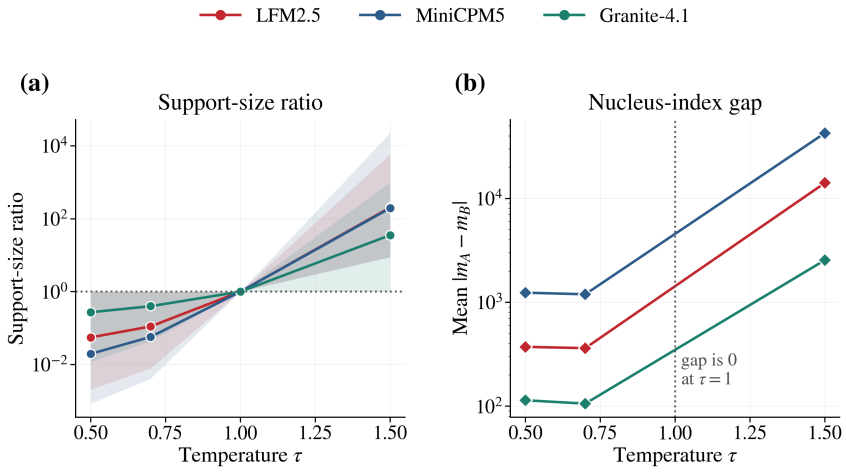

Figure 3: Support-size ratio and nucleus-index gap under temperature at $p_0 = 0.95$ over the full WikiText-103 test split, as a function of τ, for the three models. (a) Median ratio of the tempered-nucleus support size to the base-nucleus support size on a logarithmic axis, with the shaded band giving the 5th to 95th percentile range and the dotted line marking a ratio of 1. (b) Mean nucleus-index gap $|m_A - m_B|$ on a logarithmic axis, with $\tau = 1$ omitted because the gap is exactly zero there.