Varentropy Sampling

Anonymous ACL submission

Abstract

Truncation-based decoding methods reduce degeneration in open-ended language generation, but rely on first-order distribution summaries that collapse distributions with substantially different surprisal dispersion. Achieving an effective trade-off between repetition and diversity whilst maintaining natural frequency statistics and distributional fidelity in language generation is thus an open challenge. We introduce *Varentropy Sampling*, a decoding strategy that ranks candidate tokens using the product of probability and surprisal deviation from the distribution mean, preferentially retaining tokens whose surprisal contribution is high relative to the distribution's variance. We evaluate against established decoding baselines on WikiText-103, OpenWebText, and WritingPrompts using Qwen2-7B, Llama3-8B, and GPT2-XL. Varentropy ($\tau = 0.95$) achieves optimal or competitive repetition, the highest corpus-level diversity and Zipf slopes closest to human-like -1 scaling across nearly all evaluated settings. While typical sampling achieves lower repetition on creative text through aggressive support collapse (KL > 1.0), Varentropy sampling maintains substantially lower distributional distortion across all settings, with KL divergence an order of magnitude below aggressive truncation methods. These results suggest surprisal-dispersion-based selection provides an alternative decoding regime that maintains lexical breadth and distributional fidelity while remaining robust across architectures and domains.

1 Introduction

Open-ended natural language generation requires language models to produce coherent, diverse continuations from short prompts. Auto-regressive language models define a probability distribution over the next token at each step, but sampling directly from this distribution is known to produce incoherent or degenerate output (Holtzman et al., 2020; Meister et al., 2023; Welleck et al., 2020; Nadeem et al., 2020; Su et al., 2022; Finlayson et al., 2024). As such, decoding methods typically truncate the distribution before sampling. Such methods include top-k (Fan et al., 2018), nucleus or top-p (Holtzman et al., 2020), typical (Meister et al., 2023) and η-sampling (Hewitt et al., 2022).

These methods largely rely on first-order summaries of the next-token distribution, such as token rank, cumulative probability mass, or entropy. While effective, these summaries often collapse distributions with substantially different surprisal dispersion to similar scalar values. In particular, two distributions may exhibit identical entropy while differing markedly in the dispersion of token surprisals.

To capture this distinction, we consider the variance of surprisal under the model distribution, which we refer to as *Varentropy*. Varentropy can be defined as:

$$\text{Varentropy}(p) = \sum_{v} p(v)(s(v) - H)^2 \quad (1)$$

where $p(v)$ denotes the probability assigned to token v, $s(v) = -\log p(v)$ is the surprisal of token v, and $H = \sum_{v} p(v)s(v)$ is the Shannon entropy of the next-token distribution.

Whereas entropy measures the expected surprisal of the distribution, varentropy measures the dispersion of surprisal values around that expectation. High varentropy indicates greater heterogeneity in token surprisals, while low varentropy corresponds to more homogeneous distributions.

We hypothesize that selecting tokens by their varentropy contribution produces a support set better aligned with the distribution's structural heterogeneity than rank or entropy-based criteria. We propose *Varentropy sampling*, a truncation strategy that ranks tokens according to their contribution to the total varentropy:

$$score(v) = p(v)(s(v) - H)^2 \quad (2)$$

This score corresponds exactly to the per-token contribution to the total varentropy.

Unlike typical sampling (Meister et al., 2023), which preferentially retains tokens whose surprisals lie near the entropy of the distribution, our method emphasizes tokens contributing most strongly to surprisal variance while still maintaining substantial probability mass.

The probability weight $p(v)$ ensures that tokens with large distance from H but negligible probability mass do not dominate the ranking: a token must be both probable and contribute substantially to surprisal variance to score highly. We accumulate probability mass in descending score order up to threshold $\tau \in (0, 1)$ (analogous to p in nucleus sampling) and discard the remainder. The retained distribution is subsequently re-normalized before sampling.

We evaluate Varentropy sampling on WikiText-103 (Merity et al., 2017), OpenWebText (Gokaslan and Cohen, 2019), and WritingPrompts (Fan et al., 2018; Euclaise, 2018) using Qwen2-7B (Yang et al., 2024) to assess cross-domain performance, and extend evaluation to Llama-3-8B (Grattafiori et al., 2024) and GPT2-XL (Radford et al., 2019) to assess cross-family generalization. Our main findings are as follows:

- Varentropy sampling ($\tau = 0.95$) achieves lowest repetition and highest corpus-level diversity among all truncation methods across three domains and three model families; typical sampling ($\tau = 0.2$) achieves lower repetition on WritingPrompts but at $16\times$ higher KL divergence.

- Varentropy sampling maintains substantially broader support sets than most methods, retaining upwards of 2,000 tokens and up to $\approx$7,000 compared to $\leq$40 for methods including top-k, min-p and typical sampling ($\tau = 0.2$) while achieving lower or equivalent repetition.

- Varentropy sampling achieves the lowest KL divergence among all truncation methods maintaining distributional fidelity despite retaining substantially larger support sets with Zipf slopes closest to -1 scaling across nearly all evaluated settings.

We release code for all experiments.[1]

2 Related Work

2.1 Truncation-Based Sampling

Truncation-based methods address the degeneration problem in open-ended text generation by restricting the sampling vocabulary at each step to a subset of tokens. The resulting sampling distribution is defined by:

$$q(v) = \begin{cases} \dfrac{p(v)}{\sum_{u \in \mathcal{S}} p(u)} & v \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $p(v)$ is the model's original next-token distribution (softmax over logits), and $\mathcal{S} \subseteq \mathcal{V}$ is a support set over the vocabulary $\mathcal{V}$. Truncation methods define this retained token set at each decoding step and renormalize the distribution over that set before sampling. Methods differ primarily in how $\mathcal{S}$ is constructed.

Top-k sampling (Fan et al., 2018) retains the k highest-probability tokens regardless of the shape of the distribution, applying the same support size whether the distribution is highly peaked or nearly uniform. Its limitations are well documented by (Holtzman et al., 2020), who propose **nucleus (top-p) sampling** to define support adaptively. Nucleus sampling retains the smallest set of tokens whose cumulative probability exceeds a threshold p where tokens are accumulated in descending order of their probability $P(x \mid x_{<t})$. This allows the effective vocabulary to expand and contract with the distribution's shape. Nucleus sampling remains the dominant baseline in the open-ended generation literature.

Min-p sampling (Minh et al., 2025) retains tokens satisfying $p(v) \geq \alpha \cdot p_{\max}$, where $p_{\max}$ is the probability of the most likely token and α is a relative threshold. This scale-invariant truncation rule performs well at low temperatures where the distribution is highly peaked, but as temperature increases and $p_{\max}$ decreases, the effective threshold may become permissive, potentially admitting a large number of low-probability tokens.

η sampling (Hewitt et al., 2022) defines an entropy-adaptive probability threshold

$$\varepsilon(H) = \min\left(\eta,\ \sqrt{\eta}\, e^{-H}\right)$$

[1] URL redacted for review.

and retains all tokens satisfying $p(v) \geq \varepsilon(H)$. Unlike nucleus sampling, which constrains cumulative probability mass, η-sampling imposes a lower bound on individual token probability. As entropy increases, the threshold decreases adaptively, allowing the effective support size to expand in uncertain contexts while aggressively truncating low-probability tails in more confident distributions.

Locally typical sampling (Meister et al., 2023) is motivated by the observation that high-probability sequences tend to concentrate around the conditional entropy of the distribution. At each decoding step, tokens are ranked according to their deviation from the entropy, $|s(v) - H|$, where $s(v) = -\log p(v)$ is token surprisal and H is the entropy of the next-token distribution. Probability mass is then accumulated in ascending order up to threshold τ, preferentially retaining tokens whose surprisals are locally typical under the model distribution.

None of these methods consider the dispersion of surprisal values around the mean.

2.2 Stateful and Adaptive Sampling

The methods described in Section 2.1 are *stateless* — the truncation decision at each step depends only on the current next-token distribution. Stateful methods such as Mirostat (Basu et al., 2021) maintain running state across steps, using feedback control to stabilize observed perplexity of generated text precluding direct comparison under our evaluation protocol. Zhou et al. (2025) propose an adaptive framework requiring offline corpus preprocessing, tying it to a fixed target domain while BAT sampling (Finlayson et al., 2024) derives adaptive truncation thresholds via statistical estimation aimed at conservative uncertainty control. Unlike these approaches, Varentropy sampling is stateless, requires no corpus statistics or feedback control, and operates from a single hyperparameter, making it a direct drop-in replacement for existing truncation methods.

3 Experimental Setup

Varentropy sampling is assessed against all truncation-based sampling methods outlined in Section 2, including top-k, nucleus sampling, min-p, η-sampling, and typical sampling, at p and τ values recommended in literature. Temperature sampling and beam search are also included as baselines. We restrict our comparison to stateless truncation methods, which make support decisions using only the current next-token distribution. Stateful methods such as Mirostat (Basu et al., 2021), which maintain running state across decoding steps, are excluded for methodological consistency.

To ensure portability and reproducibility, all decoding methods were implemented as self-contained stateless PyTorch tensor functions rather than relying directly on framework-specific generation APIs. Implementations were validated using automated unit tests comparing retained support sets and post-truncation probability distributions against the Hugging Face `transformers` (Wolf et al., 2020) reference implementations within floating-point tolerance under diverse synthetic logit distributions.

3.1 Models and Datasets

We evaluate Varentropy sampling across multiple model families and text domains to assess both architectural and domain-level generalization. Experiments span encyclopedic text, open-domain web prose, and long-form creative fiction using decoder-only transformer language models with distinct training corpora and tokenizer designs.

All models were evaluated in their pretrained base form without task-specific fine-tuning or instruction tuning. This isolates the effect of the decoding strategy itself, avoiding confounds introduced by supervised adaptation or preference optimization.

We evaluate on three open-ended generation benchmarks spanning distinct linguistic regimes.

- **WikiText-103** (Merity et al., 2017): encyclopedic Wikipedia text with relatively low lexical entropy and strong local coherence, emphasizing fluency and repetition control.
- **OpenWebText** (Gokaslan and Cohen, 2019): heterogeneous web-scale prose with broader stylistic variability and noisier discourse structure, providing a diverse open-domain setting.
- **WritingPrompts** (Fan et al., 2018; Euclaise, 2018): long-form creative fiction with higher uncertainty and long-range narrative dependencies, making it particularly sensitive to degeneration and diversity failures.

For each dataset, prompts were constructed by tokenizing documents using the native tokenizer of the generating model and extracting the first 32

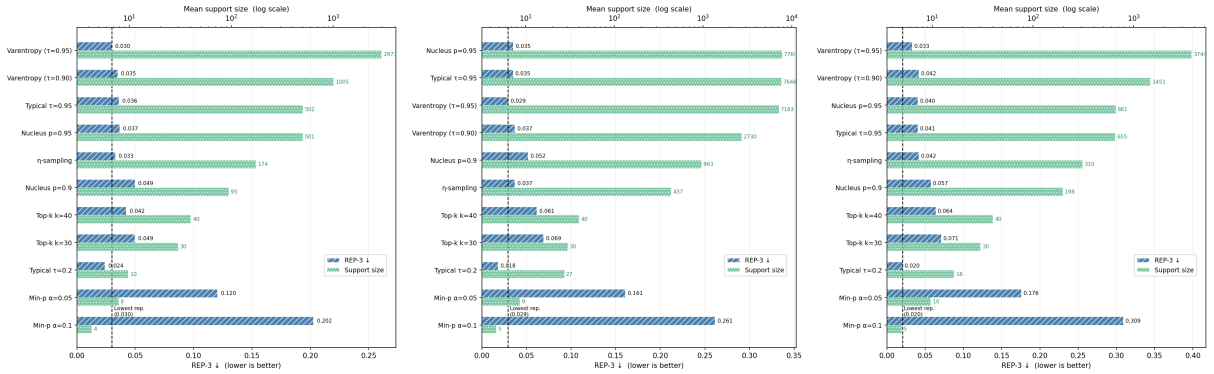

Figure 1: REP v support across three model families on WritingPrompts. Visualizing support size against repetition illustrates the tendency of best performing methods to achieve a better tradeoff between maximizing support size and minimizing repetition. Some methods that appear to perform well in minimizing repetition in Table 1 achieve so only by drastically reducing support. Notably, support sizes differ by several orders of magnitude across methods, motivating the logarithmic support axis. Left: Qwen2-7B, middle: Llama3-8B, right: GPT2-XL

tokens as context. The subsequent 256 tokens were retained as reference continuations for evaluation metrics such as MAUVE and perplexity. Documents shorter than the required prompt and reference length were discarded. Since OpenWebText does not define a canonical evaluation split, held-out documents were sampled directly from the corpus using a fixed random seed. For WikiText-103, we use the training split when large-scale evaluation ($N = 5000$) is required, since the test split contains insufficient long-form examples after filtering.

3.2 Hyperparameters

All experiments used 32 prompt tokens (Vakilian et al., 2025; Holtzman et al., 2020) and a maximum generation length of 256 tokens (Meister et al., 2023; Su et al., 2022). All sampling methods were evaluated at $T = 1.0$, which is standard for open-ended generation benchmarks and allows the model's natural distribution to be expressed without pre-scaling, ensuring truncation methods are compared on equal footing (Holtzman et al., 2020; Meister et al., 2023). Beam search with beam width $k = 5$ is included as a deterministic baseline representing maximum fluency at the cost of diversity (Meister et al., 2023; Su et al., 2022), rather than as a competitive truncation method. All experiments were run with $N = 5{,}000$ samples to ensure reliable MAUVE scoring (Pillutla et al., 2021), with a fixed random seed of 1234, consistent with prior work at this sample size (Meister et al., 2023). A single seed was used since variance across seeds is negligible at that scale (Madaan et al., 2025). All experiments were conducted on a single NVIDIA A100-PCIE-40GB GPU.

3.3 Evaluation Metrics

We evaluate all methods across six automatic quality and diversity metrics. We measure repetition (REP↓) using the rep-3 metric of Welleck et al. (2020), which computes the proportion of repeated trigrams within a generated sequence, averaged over all generations. Lower values indicate less repetitive text.

Following Meister et al. (2023), we evaluate lexical diversity using distinct-n (Li et al., 2016) averaged over $n \in \{2, 3, 4\}$. Sequence-level diversity is defined as the average fraction of unique versus total n-grams within each generated sequence:

$$D_s = \frac{\#\text{ unique } n\text{-grams in sequence}}{\#n\text{-grams in sequence}}. \quad (4)$$

Corpus-level diversity is computed analogously over the full generated corpus:

$$D_c = \frac{\#\text{ unique } n\text{-grams in corpus}}{\#n\text{-grams in corpus}}. \quad (5)$$

We report both metrics in the main evaluation tables, but use sequence-level diversity for paired bootstrap significance testing since prompts constitute the independent resampling units, whereas corpus-level diversity is a non-decomposable global statistic. Differences between Varentropy sampling and baseline methods are evaluated using paired bootstrap resampling ($B = 1000$) over the $N = 5000$ generated sequences. We report 95% confidence intervals for both REP and sequence-level diversity differences; intervals excluding zero

are considered statistically significant. Full bootstrap results are provided in Appendix B.

We evaluate fluency using perplexity under GPT-Neo-1.3B (Black et al., 2021), an external scorer from a different model family ($\mathrm{PPL}(g) \downarrow$), and under the generating model itself ($\mathrm{PPL}(i) \downarrow$), which measures consistency with the model's own distribution independently of scorer choice.

MAUVE↑ (Pillutla et al., 2021) measures distributional similarity between generated and reference text in a pretrained embedding space, with higher values indicating greater similarity to human text. To ensure reliable estimation, we compute MAUVE over all $N = 5{,}000$ sequences following Pillutla et al. (2021).

Following Meister et al. (2023), we evaluate lexical frequency statistics using the Zipf slope of the generated token distribution. Let $f(r)$ denote the frequency of the token with rank r in the generated corpus. Zipf's law states that natural language approximately follows a power-law relationship $f(r) \propto r^{-s}$, which becomes linear in log-log space as $\log f(r) = -s \log r + c$, where s is the Zipf exponent and c is a constant. We report the fitted regression slope directly (rather than its absolute value), such that values closer to -1 correspond to more natural Zipfian scaling behavior (Zipf, 1949; Piantadosi, 2014). Steeper slopes (< -1) indicate lexical concentration and over-reliance on high-frequency tokens, while flatter slopes (> -1) indicate comparatively diffuse lexical usage and increased utilization of lower-frequency tokens.

We report the per-step $\mathrm{KL}(q \| p)$ between the truncated distribution q and the model distribution p, averaged over decoding steps, as a measure of distributional distortion introduced by truncation. Full results are in Appendix A.

4 Results and Analysis

4.1 Quality and Diversity Results

Tables 1 and 2 show highly consistent decoding behavior across both domains and model families. Varentropy sampling ($\tau = 0.95$) achieves the highest corpus-level diversity and lowest or near-lowest repetition among all truncation methods across WikiText-103, WebText, and WritingPrompts on Qwen2-7B, and this pattern holds for Llama-3-8B and GPT2-XL. On WebText, Varentropy ($\tau = 0.9$) additionally achieves the highest MAUVE score overall (0.9444), and boasts competitive MAUVE scores on the other model-domain results.

The one consistent exception is typical sampling ($\tau = 0.2$), which achieves significantly lower repetition and higher sequence-level diversity on WritingPrompts across all three model families under paired bootstrap testing. However, as shown in Section 4.3, these gains come at substantial distributional cost. On Qwen2-7B WritingPrompts, Typical ($\tau = 0.2$) achieves $\mathrm{REP} = 0.0239$ and $D_s = 0.9619$, but with lower corpus diversity ($D_c = 0.6298$), a steeper Zipf slope (-1.1412), and $\mathrm{KL} = 1.1108$ — more than $100\times$ the distortion of Varentropy ($\tau = 0.95$) ($\mathrm{KL} = 0.0095$).

Perplexity results reflect the fundamental trade-off between distributional concentration and lexical diversity. Beam search and aggressive truncation methods achieve very low $\mathrm{PPL}(i)$ and $\mathrm{PPL}(g)$ by concentrating on high-likelihood continuations at the cost of diversity. Unrestricted temperature sampling ($\tau = 1.0$) achieves strong diversity but incurs extremely high perplexity across all model families. Varentropy sampling consistently occupies an intermediate regime, preserving broad lexical coverage while avoiding the perplexity inflation of unconstrained decoding.

Qualitative inspection of generated samples (Appendix C) further illustrates these patterns. Beam search exhibits catastrophic repetition across all model families while Nucleus ($p = 0.95$), Typical ($\tau = 0.95$), and η-sampling produce near-identical continuations on several prompts, consistent with their convergent support sizes and KL divergence values. Typical ($\tau = 0.2$) produces stylistically distinct output but frequently shifts register mid-generation. Varentropy ($\tau = 0.95$) produces coherent, varied continuations across all domains, with notably specific content on WikiText-103 (naming real composers and studios) and contextually grounded dialogue on WritingPrompts.

4.2 Support Size and Degeneration

Figure 1 compares mean support size against REP across all models on WritingPrompts. Methods achieving low repetition through aggressive support collapse (typical ($\tau = 0.2$) and min-p) retain fewer than 10–30 tokens per step on average, whereas Varentropy ($\tau = 0.95$) retains 2,973 tokens on Qwen2-7B, 7,163 on Llama-3-8B, and 3,744 on GPT2-XL while still achieving the lowest or near-lowest repetition across all methods. Nucleus ($p = 0.95$) and Typical ($\tau = 0.95$) retain comparable supports ($\sim$500–7,600 tokens) but achieve higher repetition, confirming that large sup-

	Method	MAUVE(↑)	D_c (↑)	D_s (↑)	REP(↓)	Zipf	PPL(i)	PPL (g)
Wikitext-103	*Reference*	–	*0.6027*	*0.9142*	*0.0683*	*-1.2820*	–	*24.6988*
	Beam ($k = 5$)	0.0524	0.2196	0.4193	0.5747	0.8261	2.3814	5.3857
	Temperature ($\tau = 0.5$)	0.2977	0.3877	0.6905	0.2958	-1.2317	2.6097	6.4538
	Temperature ($\tau = 1.0$)	0.4606	0.6398	0.9011	0.0854	-1.0284	15.2204	31.8264
	Top-k ($k = 30$)	0.4640	0.5567	0.8627	0.1214	-1.1218	5.4450	13.5063
	Top-k ($k = 40$)	**0.4835**	0.5660	0.8685	0.1157	-1.1150	5.8101	14.3397
	Nucleus ($p = 0.9$)	0.4588	0.5562	0.8577	0.1268	-1.1111	5.6250	13.7118
	Nucleus ($p = 0.95$)	0.4462	0.5878	0.8743	0.1110	-1.0796	7.3599	17.3172
	Typical ($\tau = 0.2$)	0.3689	0.5797	0.8800	0.1063	**-1.0600**	7.3238	17.0399
	Typical ($\tau = 0.95$)	0.4375	0.5891	0.8749	0.1103	-1.0785	7.4029	17.3801
	η-sampling ($\eta = 0.0003$)	0.4577	0.5946	0.8791	0.1061	-1.0864	7.5116	17.8176
	Min-p ($p = 0.05$)	0.3798	0.5010	0.8225	0.1610	-1.1491	3.8364	9.8246
	Min-p ($p = 0.1$)	0.4128	0.4656	0.7891	0.1947	-1.1762	3.2947	8.4901
	Varentropy ($\tau = 0.9$)	0.4562	0.6011	0.8802	0.1051	-1.0746	8.4943	19.6584
	Varentropy ($\tau = 0.95$)	0.4624	**0.6145**	**0.8868**	**0.0991**	-1.0623	10.2987	**23.0448**
WebText	*Reference*	–	*0.7073*	*0.9531*	*0.0349*	*-1.0421*	–	*16.3177*
	Beam ($k = 5$)	0.1080	0.2973	0.4770	0.5154	-1.2334	2.2222	3.286
	Temperature ($\tau = 0.5$)	0.6350	0.4746	0.9616	0.2304	-1.1762	2.9787	4.7077
	Temperature ($\tau = 1.0$)	0.9233	0.7181	0.7536	0.0271	-1.0072	20.3273	28.7431
	Top-k ($k = 30$)	0.9345	0.6455	0.9384	0.0458	-1.0873	7.6215	11.4342
	Top-k ($k = 40$)	0.9371	0.6534	0.9433	0.0415	-1.0812	8.1813	12.2276
	Nucleus ($p = 0.9$)	0.9400	0.6532	0.9388	0.0461	-1.0782	8.1980	12.2834
	Nucleus ($p = 0.95$)	0.9380	0.6781	0.9497	0.0362	-1.0569	10.8031	16.0010
	Typical ($\tau = 0.2$)	0.9276	0.6783	**0.9572** ‡	0.0309‡	-1.0724	11.0600	16.2312
	Typical ($\tau = 0.95$)	0.9380	0.6781	0.9502	0.0362	-1.0565	10.8031	16.0010
	η-sampling ($\eta = 0.0003$)	0.9419	0.6824	0.9523	0.0342	-1.0566	11.0819	**16.3643**
	Min-p ($p = 0.05$)	0.9059	0.5957	0.9050	0.0770	-1.1134	5.1701	7.8528
	Min-p ($p = 0.1$)	0.8732	0.5585	0.8673	0.1145	-1.1345	4.1769	6.4585
	Varentropy ($\tau = 0.9$)	**0.9444**	0.6893	0.9514	0.0355	-1.0435	12.8047	18.9442
	Varentropy ($\tau = 0.95$)	0.9343	**0.7011**	0.9568	**0.0308**	**-1.0300**	14.7670	21.6293
WritingPrompts	*Reference*	–	*0.6265*	*0.9575*	*0.0263*	*-1.1267*	–	*25.4780*
	Beam ($k = 5$)	0.0350	0.1823	0.3399	0.6541	-1.3810	2.2744	2.5249
	Temperature ($\tau = 0.5$)	0.1466	0.3070	0.5706	0.4170	-1.3263	2.4544	3.6350
	Temperature ($\tau = 1.0$)	0.6925	0.6907	0.9617	0.0251	-1.0162	28.7492	36.6093
	Top-k ($k = 30$)	0.7240	0.5639	0.9305	0.0495	-1.1696	8.4329	11.5972
	Top-k ($k = 40$)	0.7183	0.5777	0.9384	0.0422	-1.1617	9.2259	12.6500
	Nucleus ($p = 0.9$)	0.7195	0.5834	0.9317	0.0495	-1.1359	9.7866	13.3268
	Nucleus ($p = 0.95$)	0.7518	0.6257	0.9467	0.0366	-1.0985	13.7325	18.4403
	Typical ($\tau = 0.2$)	0.6701	0.6298	**0.9619** †	**0.0239**†	-1.1412	14.6069	19.9401
	Typical ($\tau = 0.95$)	0.7286	0.6267	0.9474	0.0360	-1.0964	13.9308	18.6495
	η-sampling ($\eta = 0.0003$)	0.7517	0.6298	0.9508	0.0327	-1.1072	13.5778	18.2857
	Min-p ($p = 0.05$)	0.5949	0.4798	0.8582	0.1203	-1.2238	5.0943	7.2190
	Min-p ($p = 0.1$)	0.4327	0.4191	0.7782	0.2025	-1.2569	3.7927	5.5747
	Varentropy ($\tau = 0.9$)	**0.7524**	0.6440	0.9494	0.0348	-1.0774	16.2140	21.8073
	Varentropy ($\tau = 0.95$)	0.7300	**0.6624**	0.9553	0.0298	**-1.0588**	19.7565	**26.2349**

Table 1: Quality and Diversity Metrics on Qwen2-7B across three domains. Bolded values indicate the best-performing truncation method: highest MAUVE, highest D_c (corpus-level diversity), D_s (per-sequence diversity), lowest REP, Zipf closest to -1 and $PPL(g)$ closest to reference. Temperature sampling and beam search are non-truncation baselines excluded from bolding. † indicates the marked method has significantly lower REP or higher D_s than Varentropy ($\tau = 0.95$). ‡ indicates no statistically significant difference from Varentropy ($\tau = 0.95$). All other truncation methods are significantly outperformed by Varentropy ($\tau = 0.95$) on both metrics. Full bootstrap tables are provided in Appendix B.

	Method	MAUVE(↑)	D_c (↑)	D_s (↑)	REP(↓)	Zipf	PPL(i)	PPL (g)
Meta-Llama3-8B	*Reference*	–	*0.6285*	*0.9578*	*0.0260*	*-1.1268*	–	*25.4504*
	Beam ($k = 5$)	0.0225	0.1595	0.2858	0.7098	-1.3548	65.1501	2.9779
	Temperature ($\tau = 0.5$)	0.0796	0.2748	0.4975	0.4921	-1.3418	2.6371	3.2448
	Temperature ($\tau = 1.0$)	0.5409	0.7282	0.9660	0.0226	-0.9495	125.5927	82.1317
	Top-k ($k = 30$)	0.6960	0.5628	0.9114	0.0691	-1.1611	10.2560	12.6108
	Top-k ($k = 40$)	**0.6993**	0.5787	0.9199	0.0613	-1.1456	11.2826	13.8309
	Nucleus ($p = 0.9$)	0.6837	0.6121	0.9311	0.0517	-1.0873	28.2651	**24.5562**
	Nucleus ($p = 0.95$)	0.6490	0.6600	0.9503	0.0348	-1.0319	65.8494	42.6860
	Typical ($\tau = 0.2$)	0.6527	0.6650	**0.9702**†	**0.0180**†	-1.0974	54.7165	45.4688
	Typical ($\tau = 0.95$)	0.6617	0.6601	0.9504	0.0347	-1.0319	63.7917	42.5283
	η-sampling ($\eta = 0.0003$)	0.6863	0.6530	0.9481	0.0369	-1.0692	29.0119	29.0306
	Min-p ($p = 0.05$)	0.5114	0.4691	0.8192	0.1606	-1.2205	5.9902	7.4077
	Min-p ($p = 0.1$)	0.3092	0.3988	0.7213	0.2610	-1.2612	4.2881	5.3814
	Varentropy ($\tau = 0.9$)	0.6668	0.6763	0.9490	0.0368	-1.0196	55.7683	44.0850
	Varentropy ($\tau = 0.95$)	0.6157	**0.7001**	0.9578	0.0294	**-0.9865**	86.2497	59.2431
GPT2-XL	*Reference*	–	*0.5933*	*0.9185*	*0.0678*	*-1.1199*	–	*25.7712*
	Beam ($k = 5$)	0.0051	0.0558	0.1088	0.8905	-1.4820	1.724	1.734
	Temperature ($\tau = 0.5$)	0.0338	0.1924	0.3814	0.6118	-1.3885	2.4310	2.6691
	Temperature ($\tau = 1.0$)	0.5127	0.7139	0.9631	0.0265	-0.9764	42.0721	58.5686
	Top-k ($k = 30$)	0.5057	0.5259	0.9109	0.0709	-1.1692	9.4178	11.5509
	Top-k ($k = 40$)	0.5638	0.5429	0.9189	0.0636	-1.1581	10.3131	12.7346
	Nucleus ($p = 0.9$)	0.5358	0.5929	0.9274	0.0574	-1.0946	13.7453	17.5533
	Nucleus ($p = 0.95$)	0.5569	0.6453	0.9463	0.0404	-1.0479	20.4324	27.0102
	Typical ($\tau = 0.2$)	**0.5900**	0.6436	**0.9681**†	**0.0203**†	-1.1275	21.3708	27.3216
	Typical ($\tau = 0.95$)	0.5681	0.6448	0.9462	0.0405	-1.0505	20.3775	**26.9422**
	η-sampling ($\eta = 0.0003$)	0.5873	0.6318	0.9445	0.0416	-1.0657	17.9178	23.2309
	Min-p ($p = 0.05$)	0.4034	0.4238	0.8060	0.1756	-1.2236	5.9151	7.0353
	Min-p ($p = 0.1$)	0.2355	0.3374	0.6757	0.3090	-1.2745	4.2821	4.9520
	Varentropy ($\tau = 0.9$)	0.5711	0.6600	0.9456	0.0418	-1.0247	23.7805	31.7335
	Varentropy ($\tau = 0.95$)	0.5419	**0.6843**	0.9555	0.0327	**-1.0073**	30.0371	40.9146

Table 2: Quality and diversity metrics on WritingPrompts for Meta-Llama3-8B and GPT2-XL, evaluating cross-family generalization of decoding behavior. Despite substantial architectural differences, the relative tradeoffs between truncation methods remain highly consistent across models. Typical sampling ($\tau = 0.2$) achieves the lowest repetition and highest sequence-level diversity through aggressive truncation, while Varentropy sampling consistently preserves stronger corpus-level diversity and Zipf structure with substantially broader supports. Bolded values denote best-performing methods (as described in 1). † indicates statistically significant improvement over Varentropy ($\tau = 0.95$) under paired bootstrap resampling ($B = 1000$, $N = 5000$).

port alone does not imply low repetition, nor does low repetition require aggressive truncation. The support–repetition tradeoff is stable across all three model families, suggesting it reflects a general property of surprisal-dispersion-based truncation rather than model-specific behavior. This suggests Varentropy sampling selects support sets that better preserve distributional diversity while suppressing degenerative repetition.

4.3 Distributional Distortion

Figure 2 and Table 3 reveal the distributional cost of each method's repetition reduction strategy. Across all three model families, Varentropy ($\tau = 0.95$) occupies the favorable bottom-left position of the REP–KL plane (low repetition and low distortion simultaneously) a region shared only with *eta*-sampling, nucleus sampling ($p = 0.95$) and typical sampling (0.95). Typical ($\tau = 0.2$) achieves the lowest repetition but at $\mathrm{KL} > 1.0$, placing it in the top-left corner: degeneration reduction through extreme distribution collapse. Min-p achieves neither: high repetition and high KL. Top-k and nucleus cluster in an intermediate zone with moderate repetition and moderate distortion.

The KL table further shows this holds across domains on Qwen2-7B: Varentropy ($\tau = 0.95$) achieves $\mathrm{KL} = 0.023$–0.029 across WikiText-103 and WebText, consistently below η-sampling, nucleus, and top-k at comparable repetition levels. The positioning of all methods in the REP–KL plane is remarkably consistent across Qwen2-7B, Llama-3-8B, and GPT2-XL, confirming that the observed behavior generalizes across architectures.

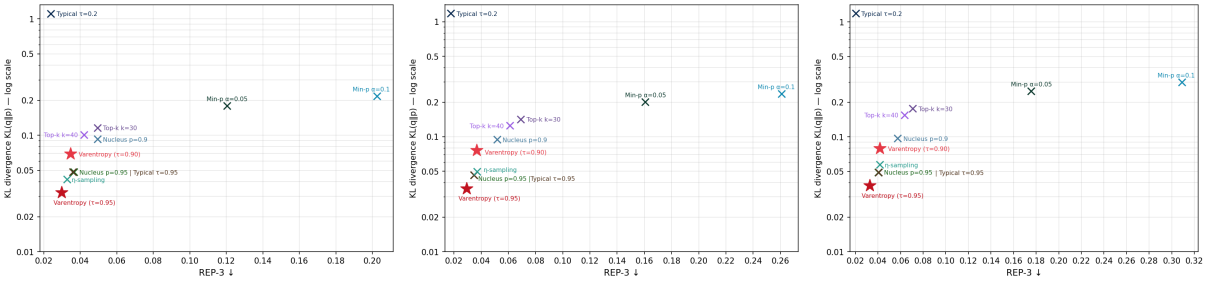

Figure 2: **Repetition versus distributional distortion on WritingPrompts across model families.** Each point shows the tradeoff between degeneration reduction (REP-3) and distortion of the original next-token distribution, measured by $\mathrm{KL}(q|p)$ between the truncated distribution q and the model distribution p. Note logarithmic KL axis. Left: Qwen2-7B, middle: Llama3-8B, right: GPT2-XL

4.4 Lexical Scaling and Zipf Structure

Prior work in quantitative linguistics has observed that natural language token frequencies approximately follow a Zipf law with exponent near -1 (Piantadosi, 2014; Zipf, 1949). Varentropy ($\tau = 0.95$) consistently yields Zipf slopes closest to -1 across all models and domains, and in several cases closer to -1 than the reference corpus itself. Methods such as min-p and top-k produce substantially steeper slopes, indicating concentration among high-frequency lexical items consistent with their support-collapse behavior illustrated by Figure 1. Unlike temperature ($\tau = 1.0$), which also flattens Zipf slopes but exhibits substantially worse degeneration, Varentropy maintains natural lexical scaling while simultaneously preserving low repetition and low KL divergence.

4.5 Threshold Ablation

Figure 3 shows the effect of τ on Qwen2-7B / WikiText-103. REP decreases and diversity increases monotonically with τ, while $\mathrm{PPL}(g)$ rises as the support broadens toward pure sampling. Low-threshold regimes ($\tau \leq 0.2$) exhibit near-greedy behavior with concentrated supports, elevated repetition, and reduced diversity. The selected value $\tau = 0.95$ lies at the knee of the REP–diversity curve, achieving strong degeneration reduction before the sharp perplexity increase associated with near-unconstrained decoding.

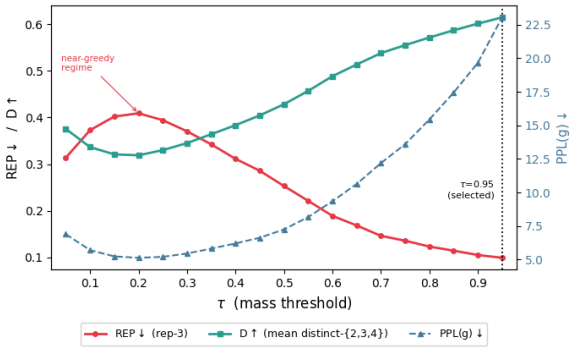

Figure 3: Varentropy sampling performance across mass thresholds $\tau \in [0.05, 0.95]$ on Qwen2-7B / WikiText-103. REP-3 (red, left axis) decreases monotonically as τ increases, while diversity D (teal, left axis) increases correspondingly. $PPL(g)$ (blue dashed, right axis) rises sharply at high τ as the support expands toward pure sampling.

5 Conclusion

We introduced Varentropy sampling, a truncation strategy that ranks tokens by their contribution to surprisal variance, retaining tokens that are both probable and structurally heterogeneous. Across three datasets and three model families, Varentropy ($\tau = 0.95$) consistently achieves the highest corpus-level diversity, Zipf slopes closest to -1, and the lowest KL divergence among truncation methods, with statistically significant improvements over all baselines except typical sampling ($\tau = 0.2$) on WritingPrompts — where typical sampling achieves lower repetition through support collapse at $100\times$ higher distributional distortion. These results demonstrate that degeneration reduction does not require aggressive truncation: surprisal-dispersion-based selection achieves competitive repetition and diversity while preserving broad supports and distributional fidelity across architectures and domains.

Acknowledgment

We used AI assistance for paraphrasing support during the writing of the manuscript, as well as for boilerplate code assistance.

Limitations

Our evaluation relies on automatic metrics; human evaluation following Hashimoto et al. (2019) would provide a more complete picture of the quality–diversity tradeoff and is left to future work.

As with prior work in this area, our generating models may have been exposed to the evaluation corpora during pretraining. However, since all methods under comparison use the same model, any such exposure affects all methods equally and does not confound relative comparisons.

Our evaluation is restricted to open-ended generation at temperature $T = 1.0$. The relative advantage of Varentropy sampling may vary at lower temperatures where distributions are more peaked and varentropy signal is reduced, and this sensitivity is not fully characterized beyond the ablation in Figure 3.

We evaluate three model families spanning 1.5B–8B parameters. Whether the observed behavior generalizes to instruction-tuned models, larger scales, or non-English languages remains an open question.

The threshold τ requires tuning and its optimal value may vary across domains and model families, as suggested by the difference in performance between $\tau = 0.90$ and $\tau = 0.95$ across datasets. Automatic threshold selection methods are left to future work.

References

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. Mirostat: A neural text decoding algorithm that directly controls perplexity. *Preprint*, arXiv:2007.14966.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Euclaise. 2018. Writingprompts dataset. https://huggingface.co/datasets/euclaise/writingprompts.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Conference of the Association for Computational Linguistics (ACL)*.

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

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *Preprint*, arXiv:1904.02792.

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

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. 2025. Quantifying variance in evaluation benchmarks.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

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

Moin Nadeem, Tianxing He, Kyunghyun Cho, and James Glass. 2020. 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, pages 334–346, Suzhou, China. Association for Computational Linguistics.

Steven Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Vala Vakilian, Sadegh Mahdavi, and Christos Thrampoulidis. 2025. How much context does natural language actually require? an analysis using LLMs as statistical oracles. In *ICML 2025 Workshop on Methods and Opportunities at Small Scale*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Yuxuan Zhou, Margret Keuper, and Mario Fritz. 2025. Balancing diversity and risk in llm sampling: How to select your method and parameter for open-ended text generation. *Preprint*, arXiv:2408.13586.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA.

Method	WikiText	WebText	WP
Temp ($\tau = 1.0$)	0.0009	0.0010	0.0008
Varentropy ($\tau = 0.95$)	0.0230	0.0286	0.0095
Eta-sampling (0.0003)	0.0267	0.0352	0.0418
Typical ($\tau = 0.95$)	0.0369	0.0455	0.0483
Nucleus ($p = 0.95$)	0.0369	0.0456	0.0483
Varentropy ($\tau = 0.9$)	0.0470	0.0625	0.0695
Nucleus ($p = 0.9$)	0.0680	0.0872	0.0928
Top-k ($k = 40$)	0.0615	0.0869	0.1013
Top-k ($k = 30$)	0.0705	0.1021	0.1161
Min-p ($p = 0.05$)	0.1211	0.1754	0.1799
Min-p ($p = 0.1$)	0.1582	0.2303	0.2164
Temp ($\tau = 0.5$)	0.1734	0.2407	0.1978
Typical ($\tau = 0.2$)	0.7479	1.0168	1.1108

Table 3: Distributional distortion of each method on each domain using Qwen2-7B. (WP stands for WritingPrompts). Temperature $\tau = 1.0$ is included as a lower bound on distortion.

A KL Divergence

Distributional distortion as measured using KL divergence is reported across all methods on each of the domain datasets on Qwen2-7B.

B Paired Bootstrap Resampling

This appendix presents the results of paired bootstrap resampling across Wikitext-103, OpenWebText and WritingPrompts on Qwen2-7B. It also provides results for GPT2-XL and Llama3-8B on WritingPrompts. Column labels are clarified as follows:

- *rep3_diff* is the difference in $REP3$ between varentropy sampling at $\tau = 0.95$ and individual baseline methods. Negative values occur when varentropy sampling at $\tau = 0.95$ has lower repetition.
- *rep3_ci_lo* is the lower bound of the 95% bootstrap confidence interval on the *rep3_diff*.
- *rep3_ci_hi* is the upper bound of the 95% bootstrap confidence interval on the *rep3_diff*.
- *div_diff* is the difference in mean distinct-2,3,4 between varentropy sampling and individual baselines. Positive values occur when varentropy has higher diversity.
- *div_ci_lo* is the lower bound of the 95% bootstrap CI on that difference.
- *div_ci_hi* is the upper bound of the 95% bootstrap CI on that difference.

Method	rep3_diff	rep3_ci_lo	rep3_ci_hi	div_diff	div_ci_lo	div_ci_hi
Aleatoric	-0.0060	-0.0088	-0.0032	0.0066	0.0039	0.0094
Top-k ($k = 30$)	-0.0222	-0.0256	-0.0188	0.0240	0.0206	0.0273
Top-k ($k = 40$)	-0.0165	-0.0199	-0.0134	0.0182	0.0151	0.0216
Nucleus ($p = 0.9$)	-0.0277	-0.0312	-0.0242	0.0291	0.0257	0.0326
Nucleus ($p = 0.95$)	-0.0118	-0.0148	-0.0088	0.0125	0.0096	0.0154
Typical ($\tau = 0.2$)	-0.0071	-0.0106	-0.0037	0.0068	0.0033	0.0102
Typical ($\tau = 0.95$)	-0.0112	-0.0141	-0.0082	0.0118	0.0089	0.0147
Min-p ($p = 0.05$)	-0.0619	-0.0659	-0.0580	0.0642	0.0604	0.0683
Min-p ($p = 0.1$)	-0.0956	-0.0996	-0.0916	0.0976	0.0937	0.1015
Eta-sampling ($\eta = 0.0003$)	-0.0070	-0.0103	-0.0037	0.0077	0.0046	0.0109

Table 4: Paired bootstrap results on Qwen2-7B WikiText-103 comparing standard truncation methods against the Varentropy Sampling ($\tau = 0.95$) baseline. Reported values denote mean paired differences in REP-3 and mean distinct-n diversity, with 95% bootstrap confidence intervals computed over prompts. Negative REP-3 differences indicate lower repetition than the Varentropy baseline, while positive diversity differences indicate higher distinct-n diversity.

Method	rep3_diff	rep3_ci_lo	rep3_ci_hi	div_diff	div_ci_lo	div_ci_hi
Aleatoric	-0.0046	-0.0060	-0.0031	0.0054	0.0040	0.0068
Top-k ($k = 30$)	-0.0149	-0.0166	-0.0133	0.0184	0.0168	0.0200
Top-k ($k = 40$)	-0.0106	-0.0121	-0.0092	0.0136	0.0121	0.0150
Nucleus ($p = 0.9$)	-0.0153	-0.0171	-0.0135	0.0181	0.0163	0.0199
Nucleus ($p = 0.95$)	-0.0059	-0.0074	-0.0044	0.0071	0.0057	0.0087
Typical ($\tau = 0.2$)	-0.0001	-0.0016	0.0014	-0.0003	-0.0018	0.0012
Typical ($\tau = 0.95$)	-0.0053	-0.0069	-0.0039	0.0066	0.0052	0.0082
Min-p ($p = 0.05$)	-0.0462	-0.0486	-0.0438	0.0518	0.0496	0.0542
Min-p ($p = 0.1$)	-0.0836	-0.0871	-0.0804	0.0895	0.0864	0.0929
Eta-sampling ($\eta = 0.0003$)	-0.0033	-0.0047	-0.0021	0.0045	0.0032	0.0058

Table 5: Paired bootstrap results on Qwen2-7B OpenWebText comparing standard truncation methods against the Varentropy Sampling ($\tau = 0.95$) baseline. Reported values denote mean paired differences in REP-3 and mean distinct-n diversity, with 95% bootstrap confidence intervals computed over prompts.

Method	rep3_diff	rep3_ci_lo	rep3_ci_hi	div_diff	div_ci_lo	div_ci_hi
Aleatoric	-0.0050	-0.0067	-0.0032	0.0059	0.0042	0.0075
Top-k ($k = 30$)	-0.0196	-0.0217	-0.0176	0.0248	0.0228	0.0269
Top-k ($k = 40$)	-0.0123	-0.0142	-0.0105	0.0169	0.0150	0.0187
Nucleus ($p = 0.9$)	-0.0196	-0.0220	-0.0174	0.0236	0.0215	0.0258
Nucleus ($p = 0.95$)	-0.0067	-0.0087	-0.0051	0.0086	0.0070	0.0105
Typical ($\tau = 0.2$)	0.0059	0.0040	0.0077	-0.0066	-0.0085	-0.0047
Typical ($\tau = 0.95$)	-0.0061	-0.0079	-0.0045	0.0079	0.0063	0.0097
Min-p ($p = 0.05$)	-0.0905	-0.0945	-0.0869	0.0971	0.0935	0.1009
Min-p ($p = 0.1$)	-0.1727	-0.1777	-0.1678	0.1771	0.1724	0.1821
Eta-sampling ($\eta = 0.0003$)	-0.0029	-0.0046	-0.0013	0.0046	0.0029	0.0062

Table 6: Paired bootstrap results on Qwen2-7B WritingPrompts comparing standard truncation methods against the Varentropy Sampling ($\tau = 0.95$) baseline. Positive REP-3 differences and negative diversity differences indicate methods outperforming the Varentropy baseline on the corresponding metric.

Method	rep3_diff	rep3_ci_lo	rep3_ci_hi	div_diff	div_ci_lo	div_ci_hi
Aleatoric	-0.0090	-0.0110	-0.0069	0.0099	0.0079	0.0120
Top-k ($k = 30$)	-0.0381	-0.0410	-0.0354	0.0446	0.0419	0.0474
Top-k ($k = 40$)	-0.0308	-0.0335	-0.0282	0.0366	0.0340	0.0393
Nucleus ($p = 0.9$)	-0.0246	-0.0272	-0.0221	0.0281	0.0256	0.0307
Nucleus ($p = 0.95$)	-0.0076	-0.0095	-0.0056	0.0092	0.0071	0.0111
Typical ($\tau = 0.2$)	0.0125	0.0103	0.0147	-0.0126	-0.0148	-0.0105
Typical ($\tau = 0.95$)	-0.0078	-0.0096	-0.0058	0.0093	0.0073	0.0112
Min-p ($p = 0.05$)	-0.1429	-0.1488	-0.1371	0.1495	0.1439	0.1552
Min-p ($p = 0.1$)	-0.2763	-0.2842	-0.2689	0.2798	0.2726	0.2875
Eta-sampling ($\eta = 0.0003$)	-0.0088	-0.0108	-0.0068	0.0110	0.0091	0.0130

Table 7: Paired bootstrap results on GPT2-XL WritingPrompts comparing standard truncation methods against the Varentropy Sampling ($\tau = 0.95$) baseline. Confidence intervals are computed using paired bootstrap resampling over prompts ($B = 1000$).

Method	rep3_diff	rep3_ci_lo	rep3_ci_hi	div_diff	div_ci_lo	div_ci_hi
Aleatoric	-0.0074	-0.0099	-0.0052	0.0088	0.0066	0.0112
Top-k ($k = 30$)	-0.0397	-0.0433	-0.0365	0.0464	0.0432	0.0498
Top-k ($k = 40$)	-0.0319	-0.0351	-0.0289	0.0379	0.0349	0.0410
Nucleus ($p = 0.9$)	-0.0223	-0.0251	-0.0194	0.0267	0.0237	0.0294
Nucleus ($p = 0.95$)	-0.0054	-0.0078	-0.0031	0.0074	0.0053	0.0098
Typical ($\tau = 0.2$)	0.0116	0.0095	0.0137	-0.0124	-0.0145	-0.0104
Typical ($\tau = 0.95$)	-0.0053	-0.0076	-0.0031	0.0074	0.0052	0.0097
Min-p ($p = 0.05$)	-0.1312	-0.1367	-0.1259	0.1386	0.1335	0.1440
Min-p ($p = 0.1$)	-0.2316	-0.2384	-0.2253	0.2365	0.2303	0.2432
Eta-sampling ($\eta = 0.0003$)	-0.0075	-0.0099	-0.0051	0.0097	0.0072	0.0121

Table 8: Paired bootstrap results on Llama-3-8B WritingPrompts comparing standard truncation methods against the Varentropy Sampling ($\tau = 0.95$) baseline. Reported values denote paired mean differences with 95% bootstrap confidence intervals over prompts.

Method	Generation
Prompt	The music was composed by Hitoshi Sakimoto , who had also worked on the previous Valkyria Chronicles games . When he originally heard about the project , he
Reference	*thought it would be a light tone similar to other Valkyria Chronicles games , but found the themes much darker than expected . An early theme he designed around his original vision of the project was rejected . He redid the main theme about seven times through the music production due to this need to reassess the game . The main theme was initially recorded using orchestra , then Sakimoto removed elements such as the guitar and bass , then...*
Beam ($k = 5$)	was not sure if he wanted to work on it , as he did not want to repeat himself . However , he eventually agreed to compose the music , as he felt that he could bring something new to the series . The soundtrack features a variety of musical styles , including orchestral , choral , and electronic music . The music has been praised for its emotional depth and its ability to capture the mood of the game...
Temperature ($\tau = 0.5$)	felt that the game was very similar to the previous games in the series , and he was not interested in working on it . However , after discussing the project with the game 's director , he decided to take on the project . The game 's music was released as a soundtrack album in 2017 . The soundtrack received positive reviews from critics , who praised its emotional depth and its ability to convey the game 's themes...
Temperature ($\tau = 1.0$)	felt that the theme and style of music that he had created on those games would not be appropriate for Valkyria Revolution , and he experimented with a variety of styles . This was mainly because the game did not fit with any notable niche in Japanese video game music – rivalries between Japanese game composers being a prevalent one – and he felt that " having something that has never been done before Assistant: Wayne Eugene BrattonSystem to work...
Top-k ($k = 30$)	felt that the theme and style of music that he had created on those games would not be appropriate for Valkyria Revolution , and he struggled in his attempts to convey the new and diverse settings of the title . He eventually created a completely different musical style from what he had done on the previous titles , drawing inspiration from Japanese folktales , the works of the 16th–century composer Taneyoshi Kobayashi , and 19th-century composer Joachim Raff . Although...
Top-k ($k = 40$)	felt that the theme and style of music that he had created on those games would not be appropriate for Valkyria Revolution , and he struggled in his attempts to convey the new and diverse settings of the title . He eventually created a completely different musical style from what he had done on the previous titles , drawing inspiration from older Japanese musical genres such as wadaiko , gagaku , and jōgaku . Valkyria Revolution received mixed to positive...
Nucleus ($p = 0.9$)	felt that the theme and style of music that he had created for Valkyria Chronicles was too similar to what the team wanted for this project , and that the design would not be sufficient to provide the same level of emotion . The team discussed various musical ideas with the director of the series , Masaaki Kawaguchi , and they eventually settled on the concept of using brass instruments to convey the "senior" aspect of the hero characters ....

Table 9: Sample generations from Qwen2-7B on WikiText-103 (part 1 of 2).

C Qualitative Samples

Tables 9–19 present sample generations from each decoding method across all five evaluated model–domain combinations. For each experiment, a single prompt is shown with the corresponding reference continuation and generations from all methods, truncated to 80 words. Prompts are drawn from fixed positions in the evaluation set (prompt index 2 for WikiText-103 and OpenWebText; prompt index 3 for WritingPrompts on Qwen2-7B; prompt index 7 for WritingPrompts on Llama-3-8B and GPT2-XL) using random seed 1234, consistent with the main experiments.

Method	Generation
Nucleus ($p = 0.95$)	felt that the theme and style of music that he had created for Valkyria Chronicles was too peculiar to use in a spinoff game . He therefore asked Akihisa Okabe and Jim Berenstain to create a soundtrack that would feel familiar to fans of the main series . He described the end result as a musical fusion between the series ' "cute" aesthetic and his classic "senpai" style . "I think we created a soundtrack that conveys the old yet...
Typical ($\tau = 0.2$)	felt it would be an easy one for him , as the theme for the original game had been easy for him to create . Sakimoto stated that his aim for the game 's soundtrack was to capture the " heroism and excitement of a military-themed story , as well as to depict the different emotional experiences of the characters through the music . " To help accomplish this , he also had his assistant composers include traditional German-style instruments...
Typical ($\tau = 0.95$)	felt that the theme and style of music that he had created for Valkyria Chronicles was too peculiar to use in a spinoff game . He therefore asked Akihisa Okabe and Jim Berenstain to create a soundtrack that would feel familiar to fans of the main series . He described the end result as a musical fusion between the series ' "cute" aesthetic and his classic "senpai" style . "I think we created a soundtrack that conveys the thoughts of...
η-sampling ($\eta = 0.0003$)	felt that the theme and style of music that he had created on those games would not be appropriate for Valkyria Revolution , and he experimented with a variety of styles . This was mainly because the game did not fit with any of his previous works , and it was difficult for him to create a suitable soundtrack . In an interview with Inverse , Sakimoto explained that he wanted to try to create a new style of music...
Min-p ($p = 0.05$)	felt that the theme of the game would be a challenge to compose for , and he was inspired by the setting and the characters . The soundtrack was released on February 15 , 2019 , in Japan . The soundtrack consists of 28 tracks that were created specifically for the game . Sakimoto noted that he tried to create a unique atmosphere for the game by incorporating elements such as the characters ' voices into the music . The...
Min-p ($p = 0.1$)	felt that the game was very similar to a previous game he had composed for , "Disgaea 2 : Cursed Memories" . He decided to compose the soundtrack in the same style as the "Disgaea" series , which is known for its use of choral arrangements and orchestral instruments . The game's score has been described as "elegant" and "soothing" , with "a sense of romance and beauty" . The music was well received by critics , who praised its...
Varentropy ($\tau = 0.90$)	felt that the theme and style of music for the game had to be a natural progression from the style of the previous games , and he experimented with a variety of styles . For the OP and ED themes , Hitoshi said he wanted a sound that was not too rock or too pop . He also did not want to have a voice singer , but instead wanted something that fit well with the theme . Therefore ,...
Varentropy ($\tau = 0.95$)	felt that the theme and style of music that he had created for Valkyria Chronicles was too peculiar to use in a spinoff game . He therefore asked Akihisa Okabe from Universal Music Japan , the owners of the Valkyria Chronicles brand , to recommend someone suitable to compose the music for the game . Okabe suggested Keisei Katō , who had also worked on the original game , and Sakimoto accepted . Sakimoto considered the spinoff game to be...

Table 10: Sample generations from Qwen2-7B on WikiText-103 (part 2 of 2).

Method	Generation
Prompt	Today, Toyota announced changes in executives' areas of responsibility, as well as personnel changes at the sub-executive managerial level. The most important change by far is
Reference	*the appointment of Akio Toyoda, the company's CEO and grandson of founder Kiichiro Toyoda, as President of a new 'EV Business Planning' department. Earlier this month, we reported–admittedly a little tongue-in-cheek–about Toyota announcing the creation of an electric vehicle division and putting only 4 engineers on the project with the goal to bring EVs to market by 2020. The move seems a lot more serious now that Akio Toyoda is leading the effort, and several other executives, managers, and engineers...*
Beam ($k = 5$)	the appointment of Akio Toyoda as the new president of Toyota Motor Corporation (TMC) and Toyota Motor Sales, U.S.A., Inc. (TMS), effective April 1, 2009. Akio Toyoda is the grandson of the company's founder, Kiichiro Toyoda, and the son of the company's current president, Katsuaki Toyoda. Akio Toyoda, 50, has been with the company since 1977, and has held a variety of positions, including president of Toyota Motor Sales, U.S.A., Inc. (TMS) from 2005 to 2007. He has also served...

Table 11: Sample generations from Qwen2-7B on OpenWebText (part 1 of 3).

Method	Generation
Temperature ($\tau = 0.5$)	the appointment of Akio Toyoda as the new president of Toyota Motor Corporation. Akio Toyoda will replace Katsuaki Watanabe, who will become the new president of Toyota Motor Sales, U.S.A., Inc. (TMS). Akio Toyoda is the grandson of the late Kiichiro Toyoda, the founder of Toyota Motor Corporation. Akio Toyoda will take office on April 1, 2009. He will be 50 years old at the time of his appointment. He has been a member of the Toyota Motor Corporation board...
Temperature ($\tau = 1.0$)	the elevation of Akio Toyoda to corporation president. Other changes were also made to address concerns over an aging executive command, but it is his ascension that is most designed to address increasing global challenges. The 55-year-old chief executive officer will succeed long time Toyota chairman and CEO Shinichi Okudaira in March next year.Fidelity and Management Co. increased its stake in Cells & Tissues Preservation, Inc. (NASDAQ:CTRO) by 50,825 shares based on the company's latest 13F filing with the U.S....
Top-k ($k = 30$)	the elevation of Akio Toyoda to chairman of the board for Toyota Motor Corporation. Other personnel changes include the promotion of Hidenori Seko to managing director and CEO of Toyota Europe GmbH; the establishment of the new Toyota Global Mobility Corp. with David Joyce as president; and a reshuffling of top management at its subsidiary Calty Design. Additionally, there have been promotions of several executives who will assume new posts and reports.I have no idea what David Joyce is going...
Top-k ($k = 40$)	the elevation of Akio Toyoda to chairman of the board for Toyota Motor Corporation. Other personnel changes include the promotion of Hidenori Seko to managing director and CEO of Toyota Europe GmbH; the establishment of the new Toyota Global Mobility Corp. with David Joyce as president; and a reshuffling of top management at its subsidiary Calty Design. Additionally, there have been promotions of multiple executives to managerial positions, including Mark D. Brincko, the recently retired CEO of Minda, who will...
Nucleus ($p = 0.9$)	the elevation of Akio Toyoda to chairman of the board, replacing the retiring Fujio Cho. Toyoda, who currently heads Toyota's domestic business operations, will take on his new responsibilities starting on August 1. In a related change, Takayuki Miyagawa, president of the Toyota Motor Corporation (TMC), will become the president of the Toyota group. The changes will take effect on October 1, 2009. The move to elevate Akio Toyoda to head of Toyota's board, which makes him chairman of the...
Nucleus ($p = 0.95$)	the elevation of Akio Toyoda to chairman of the board, replacing the retiring Fujio Cho. Toyoda, who currently heads Toyota's domestic business operations, will take charge of the company's global operations and the overall corporate direction. Toyota's U.S. VP of marketing, Jack Hollis, has also been named executive vice president for global sales. Under Toyoda's leadership, Toyota was faced with two crises: one in September 2009 with the sudden acceleration problem, and the other in March 2011 with the earthquake...
Typical ($\tau = 0.2$)	that 83-year-old Toyota chairman Kiyoshi Akashi is retiring and handing the top position over to the current chief operating officer (COO) Takeshi Uchiyamada. The former chief operating officer of Toyota Motor Sales U.S.A. (TMUS) was one of the founding fathers of Lexus and served as chief technical officer of the Toyota Group until 2004. As president and COO, Uchiyamada oversaw all sales operations, customer relations and corporate management. He has already made several key appointments at TMUS. Earlier this year,...
Typical ($\tau = 0.95$)	the elevation of Akio Toyoda to chairman of the board, replacing the retiring Fujio Cho. Toyoda, who currently heads Toyota's domestic business operations, will take charge of the company's global operations and the overall corporate direction. Toyota's U.S. VP of marketing, Jack Hollis, has also been named executive vice president for global sales. Under Toyoda's leadership, Toyota was faced with two crises: one in September 2009 with the sudden acceleration problem, and the other in March 2011 with the earthquake...
η-sampling ($\eta = 0.0003$)	the elevation of Akio Toyoda to chairman of the board for Toyota Motor Corporation. Other personnel changes include the promotion of Hidenori Ichihara to chief financial officer (COO) and Yasunaga Hara to chief technology officer. Two new positions, chief quality officer and chief procurement officer, are being filled by Shoichi Shibata and Ryuichi Tahara. Additionally, Shin Nagahara has been named chief operating officer and Yasuaki Nakanishi is now senior executive VP. These changes are effective April 1, and were announced...
Min-p ($p = 0.05$)	the elevation of Akio Toyoda to chairman of the board, replacing the retiring Fujio Cho. Toyoda, who currently heads Toyota's domestic business operations, will take charge of the company's global operations and the overall corporate direction. Toyota's U.S. sales chief Jim Lentz, who was elevated to president in 2012, will be promoted to group chief, which includes responsibility for Toyota Motor Sales, USA (TMUS), Toyota Motor North America (TMNA), Lexus International, and Toyota Motor Manufacturing Japan. He will also oversee...

Table 12: Sample generations from Qwen2-7B on OpenWebText (part 2 of 3).

Method	Generation
Min-p ($p = 0.1$)	the elevation of Akio Toyoda to the position of president, replacing the retiring Fujio Cho. Toyoda, who currently heads Toyota's domestic division, is a well-known figure in the company, having been the head of its global R&D operations before assuming the domestic position, and having played a prominent role in the company's post-recall crisis efforts. Other changes include the promotion of Tomoyuki Kawamura to head the domestic division, replacing Toyoda. Kawamura, who currently heads Toyota's global production division, has been...
Varentropy ($\tau = 0.90$)	the elevation of Akio Toyoda to chairman of the board for Toyota Motor Corporation. Other changes in executive responsibility include the following: Shigeki Tomoyama, CEO of Toyota Motor Corporation, will relinquish the President of Toyota Motor Corp. position and resume the post of President of Toyota Motor Sales, USA, Inc. In addition to the previously announced recruitment of the following executives, Toyota Motor Corporation is expecting to recruit seven additional executives who will establish the leadership of the company and...
Varentropy ($\tau = 0.95$)	the elevation of Akio Toyoda to corporation president. Other changes were also made to address concerns over an aging executive command, but it is his ascension that is most designed to address increasing global challenges. The 55-year-old chief executive officer will succeed long time Toyota chairman and CEO Shinichi Okudaira in March next year.Fidelity and Management Co. increased its stake in Toyota Motor Corp (NYSE:TM) by 9.2% during the 2nd quarter, according to its most recent 13F filing with the...

Table 13: Sample generations from Qwen2-7B on OpenWebText (part 3 of 3).

Method	Generation
Prompt	" No, no no no... " She backed up and turned the other way the second she saw me heading towards her. " Hey! " Starting to
Reference	*walk quickly now I picked up my pace until I nearly ran into her. She tried to duck behind one of the men carrying samples around the food court but instead just managed an awkward half step around him before I grabbed her shoulder. She turned towards me so I could get a clear view of her face. Now I was sure it was her. " Joanne? " Her eyes lit up, if only for a second before she turned...*
Beam ($k = 5$)	get annoyed, I called out to her. " What's your problem? " " I don't want to talk to you. " " Why not? " " I don't want to talk to you. " " Why not? " " I don't want to talk to you. " " Why not? " " I don't want to talk to you. " " Why not? " " I don't want to talk to you. " " Why not? " " I don't...
Temperature ($\tau = 0.5$)	panic, I grabbed her hand and pulled her towards me. " What's wrong? " " I'm fine. I just... I just don't want to see you. " She said, her voice shaking with tears. " Why? " I asked, confused. " I thought we were friends. " " I don't want to be friends with you. " She said, her voice breaking. " I don't want to be friends with anyone. " " What do you mean? " I asked,...
Temperature ($\tau = 1.0$)	panic some more, she turned around and tried running towards x. I was too late and x was half way across the street " Hey! Kaila! Wait! " I head but loosely grabbed her shirt and pulled her to get her attention before she jumped off the fence. But it was too late. Han, taking my broad shoulders, pulled me back to his position inside my white padded cell. When I turned to look, he was asleep on the ledge...
Top-k ($k = 30$)	panic I quickly pulled out a bottle of pills out of my back pocket. " I'm on my way to see my boyfriend! " " Well I'm gonna have to ask you to go to him somewhere else! " She said. " He isn't here!" I replied. " I'm a guest of his auntie!" " What? " She seemed confused. " Look, what is this? A quarantine? " I asked angrily. " How long does this last?" " Well I'm...
Top-k ($k = 40$)	panic I quickly pulled out a bottle of pills out of my back pocket. "Here, take these. Just... just try and calm down." She turned around once again and snatched the pill bottle out of my hands, she opened it and dropped its content in her fist. " Hey! What are you doing? " I grabbed her by her arm and forced her to look my way. " You are leaving. Please, go. That's the only way we can save...
Nucleus ($p = 0.9$)	panic I quickly pulled out a bottle of pills out of my back pocket. "Here, take these. Just... just try and calm down." She turned around once again and we were face to face, me with a bottle of pills and her with tears rolling down her face. I noticed the pills in my hand and quickly shut my mouth, hoping she hadn't seen. "Hey, they're just for pain. I got a toothache, okay?" I walked over to her, but...

Table 14: Sample generations from Qwen2-7B on WritingPrompts (part 1 of 2).

Method	Generation
Nucleus ($p = 0.95$)	panic I quickly pulled out a bottle of pills out of my back pocket. “Here, take these. Just sit here for two minutes and they’ll get rid of whatever’s bothering you... ” She froze in her spot, stopping her frantic movements and just stared at me like I was crazy. “No. I don’t want anything you have... ” “ Please. Just sit down and take these. They’ll calm you down, ” I assured her, holding the bottle out to her....
Typical ($\tau = 0.2$)	run I quickly grabbed her hand. “ Why do you have to follow me around like that? ” She huffed and crossed her arms over her chest. “ It’s my job, okay? I have to watch out for you, but not when you’re at work. You have your friends for that, remember? ” “ But...” “ Come on. It’s already past lunchtime and I want to have some coffee with you. ” “ Oh... you just want coffee... I knew...
Typical ($\tau = 0.95$)	panic I quickly pulled out a bottle of pills out of my back pocket. “Here, take these. Just sit here for two minutes and they’ll get rid of whatever’s bothering you... ” She froze in her spot, stopping her frantic movements and just stared at me like I was crazy. “No. I don’t want anything you have... ” “ Please. Just sit down and take these. They’ll calm you down, ” I assured her, holding the bottle out to her....
η-sampling ($\eta = 0.0003$)	panic I quickly pulled out a bottle of pills out of my back pocket. “Here, take these. Just sit here for two minutes and they’ll get rid of whatever’s bothering you... ” She froze in her spot, stopping her frantic movements and just stared at me like I was crazy. “No. I don’t want anything you have... ” “ Please. Just sit down and take these. They’ll calm you down, ” I assured her, holding the vial out towards her....
Min-p ($p = 0.05$)	panic I quickly pulled out a bottle of pills out of my back pocket. “ I’m on my way to see my boyfriend! ” “ Well I’m glad you’re doing so well, ” She said in a sarcastic tone. “ Now I just need to look in your bag to make sure you didn’t bring anything else that could be considered as a weapon. ” I quickly stuffed the pills in my pocket and zipped up my bag. “ Okay,...
Min-p ($p = 0.1$)	panic I took a step forward. She took two steps back. “ I’m not going to hurt you. I just want to talk to you. Please. ” I pleaded. “ No, no, no, no! ” She backed up to the door and shoved it open, nearly slamming it in my face. “ No! ” I took another step forward. She was about to run. “ Please, ” I begged. “ Let me talk to you. ” She slammed the door....
Varentropy ($\tau = 0.90$)	panic I took a step forward. “ Girls, calm down. ” Her hand on my shoulder stopped me. I couldn’t ever remember Kiera ever touching me before. I suppose that’s because I was always that guy who was always in conflict with her. “ I’m sorry, Jeff, but I don’t feel comfortable here with you alone. ” She shot me a sad look and what looked like panic flickered in her eyes. That’s when it hit me, she couldn’t handle...
Varentropy ($\tau = 0.95$)	panic I quickly had to calm myself down. I couldn’t just scare her. I reached out my hand and touched her shoulder “ Hey Layla, it’s me Marcus. I’m sorry. I just wanted to help you. I know you don’t really need me, I’m just trying to prove something to myself, but I’m really sorry for freaking you out. ” After she carefully withdrew her shoulder from under my touch, she crossed her arms and walked over to the door....

Table 15: Sample generations from Qwen2-7B on WritingPrompts (part 2 of 2).

Method	Generation
Prompt	“ Sadie! I told you not to stand under the tree in the middle of a thunderstorm! ” My mom frantically dialed 911,
Reference	*as I laid there, shocked. Literally. “ Hello?! My daughter was just struck- “ All of a sudden, something was wrong. I sat up, my fingertips crackling with this new power. Or was it the aftermath of the lighting? I couldn ’ t tell. Nor did I care. No. Something was definitely wrong, and it was nearby. “ If I take 295 south right now, I can get to Home Depot in ten minutes flat. And then I can finally...*
Beam ($k = 5$)	but it was too late. I had been struck by lightning and was now in a coma. The next thing I remember was waking up in a hospital bed. My mom was sitting next to me, holding my hand and crying. “Sadie, I’m so sorry,” she sobbed. “I should have listened to you.” “It’s okay, Mom,” I said weakly. “I’m just glad to be alive.” I spent the next few weeks in the hospital, recovering from my injuries. I was...
Temperature ($\tau = 0.5$)	as I lay unconscious on the kitchen floor, with a fractured skull and a broken arm. The doctors said I had a seizure, and my parents were told to keep me under close watch. I was so scared. When I woke up, I was in a hospital bed, and my parents were sitting next to me. My mom was crying, and my dad was holding my hand. I asked them what happened, and they told me that I had a...

Table 16: Sample generations from Meta-Llama-3-8B on WritingPrompts (part 1 of 2).

Method	Generation
Temperature ($\tau = 1.0$)	as I slowly disappeared into the ethers. She tried to call my name, several times, but my voice came out as static. As the storm clouds gave way to the shining points of a billion diamonds, I was surrounded by familiar faces. Well, Moony, I wasn't able to play cricket today!://BLANK>My mother sent him a message that validates the Certificate Attestation Service. I wished really hard."); Image/script:adbiseed Angeles flew us all (ENCRYPTED CHARACTER) _LOC/LatentEC iscreeen"/> enton message: *0">GhjC : So...
Top-k ($k = 30$)	as I slowly woke up with a groggy voice. She tried to keep her hands from quivering. She was trembling from fear, and the knowledge of how close she had come to losing me. As the phone rang, she clutched my hand with a vice grip. It was hard to move my arm without her squeezing the life out of my fingers. "Hi, I need the police… My daughter was struck by lightning!" "Alright ma'am, the paramedics are almost here,...
Top-k ($k = 40$)	as I slowly woke up with a groggy voice She: "We need an ambulance. My daughter is unconscious. " Me: "I'm not unconscious, mom. I'm just resting after a long day. " The ambulance showed up with a loud siren and my mom told the paramedics what happened. They asked me my name and some basic information about me while they took me inside the ambulance. I was too scared to talk about the details of what happened, so when...
Nucleus ($p = 0.9$)	as I slowly disappeared into the ethers. She tried to call my name, but I had already vanished. My uncle who happened to live in the same town as us, died from a freak thunderstorm. He was strolling along the footpath by the river. He was a normal, healthy, young man in his prime. My grandfather, whom I loved very much, used to tell me that Uncle David had been my great-great grandfather's favourite grandson, so of course he was...
Nucleus ($p = 0.95$)	as I slowly disappeared into the ethers. She tried to call my name, but I had already vanished. My uncle who happened to live in the area was also about to go into the house when he saw my mom franticly dialing the phone. He went over and asked her what was wrong. She explained what happened. My uncle immediately went into the tornado shelter that he had built into the backyard. He was really glad he did when it...
Typical ($\tau = 0.2$)	with one hand covering her face as her fingers moved along the numbers of her cellphone. My 10 year old self looked around for something to hold on to. After about five minutes, an ambulance and the fire department showed up, their red lights shining against the wall. They got my mom to a nearby hospital. After two days of intense medication and recovery, my mom finally recovered from the near death experience. Since that day, my mom became scared...
Typical ($\tau = 0.95$)	as I slowly disappeared into the ethers. She tried to call my name, but I had already vanished. My uncle who happened to live in the same town as us, died from a freak thunderstorm. He was strolling along the footpath by the river. He was a normal, healthy, young man in his prime. My grandfather, whom I loved very much, used to tell me that Uncle David had been my great-great grandfather's favourite grandson, so of course he was...
η-sampling ($\eta = 0.0003$)	as I slowly disappeared into the ethers. She tried to call my name, but I had already vanished. My uncle who happened to live in the same town as us, died from a freak thunderstorm. He was strolling along the footpath by the river. He was a normal, healthy, young man in his prime. My grandfather, whom I loved very much, used to tell me that Uncle David had been my great-great grandfather's favourite grandson, so of course he was...
Min-p ($p = 0.05$)	as I slowly woke up with a groggy voice. She tried to keep her cool as I was not able to recall anything. As I opened my eyes, I realized that I was laying in a hospital bed. A doctor was walking in through the door with a clipboard in their hand. I could not help but notice the "U.S. Army" on the back of their lab coat. I was not sure what had happened, so I asked, "why am...
Min-p ($p = 0.1$)	as I stood motionless with my mouth hanging open. She tried to pull me back from the tree, but my feet were firmly planted in the grass. I couldn't understand why she was so upset. It was just a tree. How could it hurt me? It was then that I heard it. A low, deep rumble from inside the tree. I felt it through the soles of my feet. The tree was talking to me. The rumble became words. "...
Varentropy ($\tau = 0.90$)	as I slowly disappeared into the ethers. She tried to call my name, but I had already vanished. My uncle who happened to live in the same town as us, died from a freak roof collapse. He was strolling along the footpath by his apartment. He was a normal yet quiet man. He owned a hardware store that he inherited from his father. He also smoked a lot of unfiltered Camels. A really small shed, built by my great grandfather,...
Varentropy ($\tau = 0.95$)	as I slowly disappeared into the ethers. She tried to call my name, but I had already vanished. My uncle who happened to live in the area was also about to go into the house when he saw my mom franticly mumbling " No, no, Sadie! " He took of after my mom and they found me lying under my mother's tree, drenched in the water from the nearby creek. My mother shed tears of joy as she rushed over...

Table 17: Sample generations from Meta-Llama-3-8B on WritingPrompts (part 2 of 2).

Method	Generation
Prompt	I open my eyes in a panic, sweat beading and then falling down my face. I look around and the sun in shining through the maroon curtains of
Reference	*my studio apartment. Everything seems to be as I left it the afternoon before, but there is a heavy, unfamiliar air in the room. I notice the throw cover I lazily hang to separate my twin size bed from the rest of the apartment has been ripped to the ground, and that the kitchen faucet rhythmically drips against the cold steel sink. I was alone last night like I always am, I think. No one has been here but me...*
Beam ($k = 5$)	my room. I look at the clock and it's 3:30am. I look at the clock again and it's 3:30am. I look at the clock again and it's 3:30am. I look at the clock again and it's 3:30am. I look at the clock again and it's 3:30am. I look at the clock again and it's 3:30am. I look at the clock again and it's 3:30am. I look at the clock again and...
Temperature ($\tau = 0.5$)	the window. I am in a small room with a bed and a desk. There is a door to the right of me and a door to the left of me. I look at my hands and see my fingers are covered in blood. I look at my feet and see the same thing. I look at my head and see a hole in it. I look at my eyes and see blood. I look at my mouth and see...
Temperature ($\tau = 1.0$)	the lounge is fading, the diamonds on the tray and the numerous flowering plants. As I sit up drowsily, I see a hand covered in hair, mussed with a dark brown mess, as my friend's is. He gives me a profuse hug and I begin to worry there's no way out.The sun shines brightly, again, and the homeless guy in white pulls up a chair next to me. The water is so still and I can see the ocean miles...
Top-k ($k = 30$)	the bedside table, the warm blankets of my brother and father. I start to laugh. I can hear my mother laughing and a child crying in the room. My mother is holding my brother, as my brother's eyes grow wide with surprise. "Why am I laughing?" I mumble in my haste. My brother is laughing as well. He's the best brother in the world. "Oh honey, I don't mind at all. It's not like you, this isn't funny." My mother...
Top-k ($k = 40$)	the lounge is blinding me. The air of the room looks so fresh and clean, and its just my imagination. I open my eyes and turn to look at my computer with a sudden sense of dread as my heart's racing rapidly. I'm sitting there, in front of the computer, on an Internet chat forum. You know, "Funny story time," I'm sitting there, typing furiously with the mouse. I'm trying to figure out which story was the funniest, when my...
Nucleus ($p = 0.9$)	the lounge is fading, the breeze seems to be beating slower and the orange band across the ceiling is dinted, I am a child playing in the forest, there is a dark shadow looming, as my mother's eyes grow wide with surprise, "Why in the world did you open your eyes? Did you get a snake by mistake?" "No, it was a crab," I shrug, picking up my brown blanket and fur cap which I have just left, and placing...
Nucleus ($p = 0.95$)	the lounge is fading, the diamonds on the tray and the numerous flowering plants. As I sit up my terrified mind is flooded with a million questions and it's only through conscious effort that I concentrate on the importance of planning. Through the haze of alcohol I remember hearing one of the many quotable drinks Dan says in the desert: drink as if your life depended on it. In light of all this misery I decide I want to try...
Typical ($\tau = 0.2$)	my bed is starting to pour. My clothes are stained, there's dirt in the pockets and in my shirt, I smell of a gym, but there's nothing there to remember that this was anything other than a peaceful day spent on my feet, catching up on email. I haven't stopped showering. The stink is really strong and it makes my hair wet and matted, making my whole body sweaty and irritated. The soap that comes in the bathroom was completely...
Typical ($\tau = 0.95$)	the lounge is fading, the diamonds on the tray and the numerous flowering plants. As I sit up my terrified mind is flooded with a million questions and it's only through conscious effort that I concentrate on the importance of planning. Through the haze of alcohol I remember hearing one of the many quotable drinks Dan says in the desert: drink as if your life depended on it. In light of all this misery I decide I want to try...
η-sampling ($\eta = 0.0003$)	the lounge is fading, the diamonds on the tray and the numerous pieces of jewellery and necklaces. I take a deep breath and turn to look at my daughter with a tired sigh. My wedding band is hooked to my finger, the old one, in perfect condition. The liquid on my forehead is the remnants of my drink, leftover from one of the many evenings out I have drank last night. The taxi driver is at a loss for words,...

Table 18: Sample generations from GPT2-XL on WritingPrompts (part 1 of 2).

Method	Generation
Min-p ($p = 0.05$)	the bedside table, the wall, the window and the ceiling. I start to realise something is wrong. I take a deep breath and turn to look at my daughter who is curled up on my chest. She's wearing a pink nightie, one that's cut to the midriff, and I think she's been crying. I pull the covers over her and she doesn't stir. "Shit," I say. "Something's not right. Something's not right." It's not a good day in the hospital....
Min-p ($p = 0.1$)	the bed. It's a warm, bright day and the window is open. It's too late to call a taxi. I sit up, still feeling dizzy, and feel my clothes slowly being pulled off. I stand, gasping. My legs feel like they've been bound to the bed and I'm being lifted into the air, like a rag doll being put in a washing machine. It's a good thing I'm not a rag doll. I can't feel the wet, heavy cloth of...
Varentropy ($\tau = 0.90$)	the lounge. He's here. He's right over here. "Hurry." I dawdle. He's standing by the pool, molly with a backpack slung over his shoulder. "Molly?" I gasp, "shit. What are you doing here?" "I'm sorry," he says, still sipping the molly. "I'd forgotten how rude you are when you're at meetings. It's not your fault, this isn't in my field. But they've given me an exam in human discourse and I'm here to talk to you about the foundations of...
Varentropy ($\tau = 0.95$)	the lounge is fading, the diamonds on the tray and the numerous flowering plants. As I sit up drowsily, I see a hand covered in hair, mussed with a dark brown mess, as my friend's is. He gives me a profuse hug and I begin to worry there's no way out.The sun shines brightly, again, and the homeless guy in white pulls up a chair next to me. The water is so still and I can see the ocean miles...

Table 19: Sample generations from GPT2-XL on WritingPrompts (part 2 of 2).