

---

# $p$ -less Sampling: A Robust Hyperparameter-Free Approach for LLM Decoding

---

Runyan Tan<sup>1,2</sup>, Shuang Wu<sup>1</sup>, Phillip Howard<sup>1</sup>

<sup>1</sup>Thoughtworks, <sup>2</sup>National University of Singapore

{runyan.tan, shuang.wu, phillip.howard}@thoughtworks.com

run@nus.edu.sg

## Abstract

Obtaining high-quality outputs from Large Language Models (LLMs) often depends upon the choice of a sampling-based decoding strategy to probabilistically choose the next token at each generation step. While a variety of such sampling methods have been proposed, their performance can be sensitive to the selection of hyperparameters which may require different settings depending upon the generation task and temperature configuration. In this work, we introduce  $p$ -less sampling: an information-theoretic approach to sampling which dynamically sets a truncation threshold at each decoding step based on the entire token probability distribution. Unlike existing methods,  $p$ -less sampling has no hyperparameters and consistently produces high-quality outputs as temperature increases. We provide theoretical perspectives on  $p$ -less sampling to ground our proposed method and conduct experiments to empirically validate its effectiveness across a range of math, logical reasoning, and creative writing tasks. Our results demonstrate how  $p$ -less sampling consistently outperforms existing sampling approaches while exhibiting much less degradation in text quality at higher temperature values.

## 1 Introduction

The increasingly impressive capabilities exhibited by Large Language Models (LLMs) in recent years have been aided by advancements in sampling-based decoding strategies which probabilistically select tokens at each generation step. In contrast to deterministic methods such as greedy decoding and beam search [Freitag and Al-Onaizan, 2017], sampling-based strategies can produce more diverse and human-like language outputs while avoiding issues such as neural text degeneration [Holtzman et al.]. Truncation-based sampling strategies (e.g., Top- $p$  [Holtzman et al.], Top- $k$  [Fan et al., 2018],  $\epsilon$ -sampling [Freitag et al., 2023], Mirostat [Basu et al., 2021], Min- $p$  [Nguyen et al., 2024]) have proven particularly effective by truncating the token probability distribution to only a subset of higher-likelihood tokens before sampling. However, the truncation thresholds produced by these existing approaches depend upon the specification of hyperparameters whose optimal values can vary depending upon the generation task and sampling temperature.

To address this deficiency, we introduce  $p$ -less sampling: a parameter-less sampling strategy grounded in information theory.  $p$ -less sampling uniquely combines several desirable properties of existing sampling approaches into a single method. In addition,  $p$ -less is distinguished from prior approaches in that it is parameterless. This eliminates the need to tune the sampling method’s hyperparameters, which are often sensitive to the generation task. We provide theoretical perspectives on  $p$ -less sampling to ground our proposed approach and conduct extensive experiments to demonstrate its strong empirical performance across five math, logical reasoning, and creative writing datasets. Our results and analyses show that  $p$ -less consistently produces among the best outputs across this wide range of tasks while maintaining high text quality even at higher temperature values.

## 2 Methodology

### 2.1 The $p$ -less Sampling Method

At every time step  $t$ , an autoregressive language model infers a distribution of the vocabulary tokens conditioned on the token sequence already generated from step 1 to step  $t - 1$ . Essentially,  $p$ -less considers information from the entire distribution; it computes the likelihood of a correct random guess given the distribution, which serves as our principled probability threshold adapted to the token distribution at every decoding step. We admit tokens into the sampling distribution whose likelihoods are at least that of the correct random guess likelihood. Furthermore, to determine which and how many tokens to sample from, the  $p$ -less threshold varies in a meaningfully opposite direction with entropy; as entropy increases, more tokens with lower probability are admitted for sampling.

Formally, let  $v \in \mathcal{V}$  denote the set of tokens from a vocabulary  $\mathcal{V}$ . At each time step  $t$ , let  $\mathcal{P}(\mathcal{S} = v)$  denote the probability that token  $v$  is sampled and  $\mathcal{P}(\mathcal{T} = v)$  denote the probability that token  $v$  is the correct (or most desirable) token in the "ground-truth" sense. Furthermore, let  $P_\theta(v \mid x_{1:t-1})$  denote the language model's predicted token distribution conditioned on the given token sequence  $x_{1:t-1}$ , where  $\theta$  are the language model parameters. Denoting the probability that the sampled token matches the ground-truth as  $L[P]$ , we have:

$$\begin{aligned} L[P] &:= \sum_{v \in \mathcal{V}} \mathcal{P}(\mathcal{S} = v \cap \mathcal{T} = v \mid x_{1:t-1}) \\ &= \sum_{v \in \mathcal{V}} \mathcal{P}(\mathcal{S} = v \mid x_{1:t-1}) \mathcal{P}(\mathcal{T} = v \mid x_{1:t-1}) \end{aligned} \quad (1)$$

since the sampling  $\mathcal{S}$  and correctness  $\mathcal{T}$  are independent events (no feedback involved). Notably, since we only have access to the predicted token distribution of the language model and no other external augmentation resources, we will take this as our best empirical estimate of the true token distribution, *i.e.*  $\mathcal{P}(\mathcal{T} = v) = P_\theta(v \mid x_{1:t-1})$ . Therefore, we have:

$$\begin{aligned} L[P] &= \sum_{v \in \mathcal{V}} \underbrace{\mathcal{P}(\mathcal{S} = v \mid x_{1:t-1})}_{=P_\theta(v \mid x_{1:t-1})} \underbrace{\mathcal{P}(\mathcal{T} = v \mid x_{1:t-1})}_{=P_\theta(v \mid x_{1:t-1})} \\ &= \sum_{v \in \mathcal{V}} P_\theta(v \mid x_{1:t-1})^2. \end{aligned} \quad (2)$$

We formalize the method as follows:

1. **Determine the threshold probability**  $L[P_\theta]$  with Eq.(2)
2. **Construct the sampling set**  $\mathcal{V}_{p\text{-less}}$  with tokens whose probabilities are at least  $L[P_\theta]$ :  

$$\mathcal{V}_{p\text{-less}} = \{ v \in \mathcal{V} : P_\theta(v \mid x_{1:t-1}) \geq L[P_\theta] \}. \quad (3)$$
3. **Sample from**  $\mathcal{V}_{p\text{-less}}$  the next token  $x_t$  according to the normalized token probabilities  $P'_\theta$ :

$$P'_\theta(x_t \mid x_{1:t-1})|_{x_t=v} = \frac{P_\theta(v \mid x_{1:t-1})}{\sum_{v' \in \mathcal{V}_{p\text{-less}}} P_\theta(v' \mid x_{1:t-1})} \quad \text{for } v \in \mathcal{V}_{p\text{-less}}. \quad (4)$$

### 2.2 The $p\text{-less}_{\text{norm}}$ Sampling Method

We further explore a variant of  $p$ -less which effectively relaxes the threshold by subtracting the likelihood of an incorrect random guess normalized to the number of correct outcomes. The resultant  $p\text{-less}_{\text{norm}}$ , denoted  $\bar{L}[P_\theta]$ , is preferable in use cases where diversity is favored over coherence. Formally, we have:

$$\bar{L}[P_\theta] := L[P_\theta] - \underbrace{\frac{1}{|\mathcal{V}| - 1}}_{\text{Normalization constant}} \times \underbrace{\sum_{u, v \in \mathcal{V}, u \neq v} P_\theta(u \mid x_{1:t-1}) P_\theta(v \mid x_{1:t-1})}_{\text{Probability of a randomly sampled and incorrect token}} \quad (5)$$

$$= \frac{|\mathcal{V}|}{|\mathcal{V}| - 1} L[P_\theta] - \frac{1}{|\mathcal{V}| - 1} \quad (6)$$

where  $\frac{1}{|\mathcal{V}| - 1}$  gives the ratio of the possible number of correct to incorrect outcomes. The derivation of Eq. (6) from (5) and additional details of  $p\text{-less}_{\text{norm}}$  are provided in Appendix B.4.

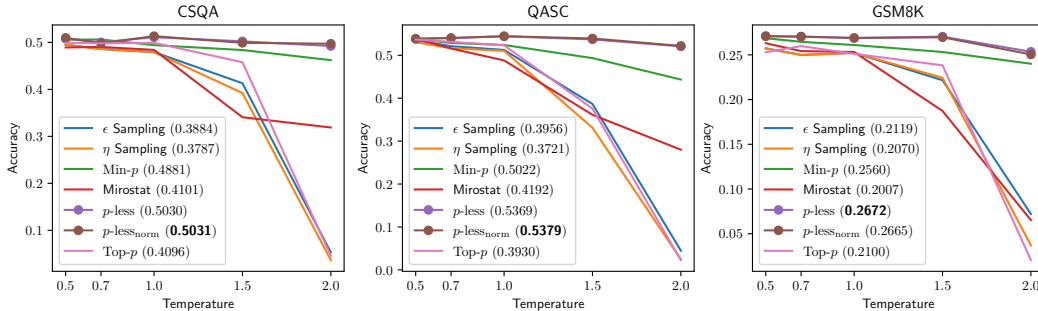
### 2.3 Theoretical Perspectives of $p$ -less

Our  $p$ -less approach can be re-interpreted in connection to established results in information theory, namely the family of Rényi entropies [Rényi, 1961]. Specifically,  $p$ -less corresponds to the exponential of the negative Rényi entropy of order 2. Intuitively, as the Rényi entropy of order 2 increases,  $p$ -less decreases. By extension of the family of Rényi entropies, we also showed  $p$ -less to vary in the opposite direction as Shannon entropy (see Appendix B.1). An alternative interpretation of  $p$ -less is that it serves as the unbiased estimator of the second moment of the distribution’s probability mass function, multiplied by the vocabulary size (see Appendix B.2). Finally, we extend  $p$ -less to a generalized  $k$ -order threshold within the formalism of Rényi entropy (Appendix B.5) and detail how it uniquely combines advantages of existing sampling methods into a single approach (Appendix B.6).

## 3 Experiments & Analyses

**Experimental setup** Our experiments were performed using Llama-2-7B (Chat) and Mistral-7B (Instruct) on two tasks, namely *math* and *logical reasoning* across the GPQA [Rein et al., 2023], GSM8K [Cobbe et al., 2021], QASC [Khot et al., 2020] & CSQA [Talmor et al., 2019] datasets, as well as *instruction following creative writing* for the Writing Prompts [Fan et al., 2018] dataset. We benchmarked our proposed sampling approaches against existing methods including Top- $p$  [Holtzman et al.], Min- $p$  [Nguyen et al., 2024],  $\epsilon$ -sampling [Freitag et al., 2023],  $\eta$ -sampling [Hewitt et al., 2022] and Mirostat [Basu et al., 2021] for temperatures ranging from 0.5 to 2.0. We measured accuracy on the math and logical reasoning datasets and computed length-controlled win rate [Dubois et al., 2024] for Writing Prompts using an automated evaluation framework [Li et al., 2023], in addition to conducting a human evaluation. Additional experiment details are provided in Appendix C.

**Math and logical reasoning results** Across the four math and logical reasoning datasets GSM8K, GPQA, QASC & CSQA,  $p$ -less and  $p$ -less<sub>norm</sub> perform superior to the other methods at temperatures 1.0 and above, and are competitive at temperatures below 1.0. To perform a fair comparison between methods across temperatures, we computed the area under the accuracy-temperature curve for each method (normalized to between 0.0 and 1.0, i.e. normalized integral), which we term AUC. For Llama2-7b, the AUCs of  $p$ -less and  $p$ -less<sub>norm</sub> outperform the other methods on all datasets, except for GPQA, where  $p$ -less still achieves the third highest AUC (see Figure 1, Table 3 and Figure 3). The results for Mistral-7b are consistent with those of Llama2-7b: the AUCs of both  $p$ -less and  $p$ -less<sub>norm</sub> outperform all other methods across every dataset (see Table 3 and Figure 4). Additional results for other hyperparameter settings are provided for Llama2-7b in Table 4 of the Appendix.



**Figure 1:** Accuracy vs. temperature curves of each method on CSQA, QASC, and GSM8k using Llama-2-7b. AUC values achieved by each method are provided in the legend (in parentheses) with the best AUC in **bold**.

**Creative writing results** We provide results for the Writing Prompts dataset using the length-controlled win rate metric [Dubois et al., 2024]. Specifically, we sample one generation per method and temperature for a subset of 100 prompts and use the response generated by default sampling (i.e. without truncation at temperature 1.0) as reference. Table 1 summarizes the results. All methods except  $p$ -less generally exhibit significant performance degradation as temperature increases. In contrast, the performance of  $p$ -less remains relatively stable and is superior to all other methods at temperatures > 1.0. This demonstrates how  $p$ -less excels in the domain of creative writing while avoiding the degradation of text quality exhibited by other methods at higher temperatures.

Model	Temperature	$\epsilon$ -sampling	$\eta$ -sampling	Min- $p$	Mirostat	Top- $p$	$p$ -less	$p$ -less <sub>norm</sub>
Llama-2-7b	1.0	<b>62.18</b>	58.76	57.48	56.94	62.07	55.08	58.74
	1.5	1.99	1.46	58.17	5.33	4.39	58.23	<b>59.58</b>
	2.0	0.00	0.00	48.94	26.88	0.00	<b>65.64</b>	59.29
Mistral-7b	1.0	60.90	59.82	66.49	62.26	65.68	<b>68.90</b>	67.49
	1.5	3.71	0.00	62.17	12.08	0.00	<b>66.97</b>	66.89
	2.0	0.00	0.00	54.11	40.33	0.00	60.32	<b>61.99</b>

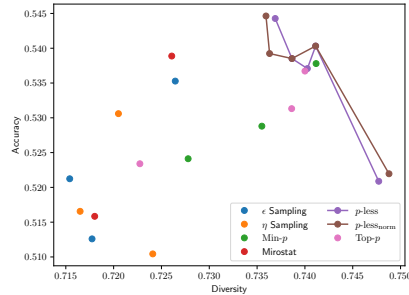
**Table 1:** Length-controlled win rate for 100 sampled prompts from the Writing Prompts dataset.

**Human evaluation** Three authors of this work evaluated 30 sampled Writing Prompts generations produced by Llama2-7b using  $p$ -less with a temperature of 2.0, under the same setting as our length-controlled win rate evaluations (i.e., pairwise comparison to default sampling). 24.1% of story pairs received unanimous agreement among the annotators; for the remaining stories, we use the majority vote to obtain a label. Overall we found that the human annotators preferred stories produced by  $p$ -less sampling 55.2% of the time. For samples in which the human annotators unanimously agreed,  $p$ -less was preferred 57.1% of the time. While this is slightly below the LLM-judged win rate for  $p$ -less (see Table 1), the directional consistency of the human and automated evaluations provides further evidence of the effectiveness of  $p$ -less sampling for creative writing.

**Diversity analysis** We compute the  $n$ -gram repetition diversity metric proposed by Su et al. [2022] for QASC; higher values indicate greater diversity. Table 2 shows that at temperatures  $\leq 1$ , all methods produce similar diversity values between 0.72-0.74. At higher temperatures,  $p$ -less and  $p$ -less<sub>norm</sub> exhibit greater diversity than min- $p$ , but lower diversity than other sampling methods. However, greater diversity at these higher temperatures leads to lower answer accuracy. Figure 2 illustrates the relationship between generation diversity and answer accuracy for Llama2-7b using sampling methods and temperature settings which achieved an overall mean accuracy  $> 0.5$  on QASC. This plot shows that  $p$ -less and  $p$ -less<sub>norm</sub> produce higher accuracy than other sampling methods at any level of generation diversity, exhibiting a pareto dominance along the diversity-accuracy frontier.

Temperature:	0.5	0.7	1.0	1.5	2.0
$\epsilon$ Sampling	0.73	0.72	0.72	0.82	0.99
$\eta$ Sampling	0.72	0.72	0.72	0.85	0.99
Min- $p$	0.74	0.74	0.73	0.71	0.73
Mirostat	0.73	0.72	0.73	0.86	0.84
Top- $p$	0.74	0.74	0.72	0.80	0.99
$p$ -less	0.74	0.74	0.74	0.74	0.75
$p$ -less <sub>norm</sub>	0.74	0.74	0.74	0.74	0.75

**Table 2:** QASC diversity by method & temperature



**Figure 2:** QASC accuracy vs. diversity

**Qualitative analysis** We found that where the other methods fail on specific math and logical reasoning examples at low temperatures,  $p$ -less and  $p$ -less<sub>norm</sub> generate correct reasoning and the correct final answer. On the other hand, when temperatures are high enough for the methods to fail,  $p$ -less and  $p$ -less<sub>norm</sub> still generate correct reasoning, including unprompted self-verification and self-correction on initial wrong answers. At high temperatures, the generations of other methods veer off-track or exhibit hallucination and degeneracy, whereas generations from  $p$ -less and  $p$ -less<sub>norm</sub> remain factual and relevant. Illustrative examples are provided in Appendix C.5. We also provide creative writing examples where  $p$ -less is unanimously preferred to default sampling by our three human evaluators and the automated evaluation framework in Appendix C.5.3.

## 4 Conclusion

We presented  $p$ -less sampling: a hyperparameter-free truncation methodology for sampling-based decoding.  $p$ -less combines several desirable properties of existing sampling methods into a single approach, resulting in higher-quality outputs across a diverse set of math, logical reasoning, and creative writing tasks. Our work highlights how grounding LLM decoding in information theory results in a principled sampling approach which is both intuitive and empirically effective.

## References

- Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- 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)*, pages 889–898, 2018.
- Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, 2023.
- 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.
- Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*, 2024.
- Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof qa benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- 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.
- Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8082–8090, 2020.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Proceedings of the NAACL-HLT*, pages 4149–4158, 2019.
- John Hewitt, Christopher D Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, 2022.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator for instruction-following language models. *GitHub repository*, 5 2023.
- Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35: 21548–21561, 2022.

## A Related Work

Text decoding in LLMs involves a truncation process where low probability tokens or those expected to disrupt coherence are excluded from the sampling pool. Top- $k$  Fan et al. [2018] restricts sampling to the  $k$  most probable tokens, which could lead to incoherent generations when the distribution is extremely uniform or peaked. Top- $p$  Holtzman et al. improves upon this by sampling from the smallest set of tokens whose cumulative probability exceeds a threshold  $p$ .  $\epsilon$ -sampling Freitag et al. [2023] proposes to truncate all tokens with probabilities below a cut-off threshold  $\epsilon$  quantity. Both Top- $p$  and  $\epsilon$ -sampling remain lacking in adapting to high-entropy conditions, which is typical when temperature is tuned up, such as for use cases where diversity is preferred. Along the  $\epsilon$ -sampling track,  $\eta$ -sampling proposes an entropy-aware variant which defines the threshold as the minimum of  $\epsilon$  and a scaled negative Shannon entropy exponential quantity. This however introduces additional hyperparameters and relies on the assumption that the entropy follows a uniform distribution baseline. A more sophisticated approach in the form of Mirostat Basu et al. [2021] assumes that the token distribution follows Zipf’s Law and dynamically adjusts the threshold to maintain a target surprisal. This dynamic feedback control involves task and model dependent tuning of the target surprisal and a learning rate. The Min- $p$  method Nguyen et al. [2024] attempts to simplify truncation through a fractional hyperparameter, which is multiplied with the modal probability to define the truncation threshold. While being more empirically robust at high temperatures, Min- $p$  remains sensitive to the choice of the fractional hyperparameter, and leverages only a single statistic from the distribution.

In contrast, the proposed  $p$ -less sampling method avoids these pitfalls by operating directly on the empirical token distribution without imposing parametric assumptions or requiring hyperparameter tuning. This approach offers a model-agnostic threshold which is robust in high-entropy regimes. The code is available at <https://github.com/ryttry/p-less>.

## B Additional Properties of $p$ -less and $p$ -less<sub>norm</sub> Sampling

### B.1 Connection to Rényi Entropies

Our  $p$ -less term can be re-interpreted in connection to established results in information theory, namely the family of Rényi entropies Rényi [1961]. The Rényi entropy of order  $\alpha$ <sup>1</sup> is defined by:

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha \quad (7)$$

In particular, the Rényi entropy of order 2 (aka the collision entropy) is given by

$$H_2(p) = -\log \sum_i p_i^2 = -\log L[P] \quad (8)$$

Since  $\log$  is a monotonically increasing function, our  $p$ -less quantity  $L[P]$  increases with decreasing collision entropy. Furthermore, we have

$$\begin{aligned} H_2(p) &= -\log L[P] \leq -\log \sum_i p_i \log p_i = H_1(p) \\ \implies L[P] &\geq \exp(-H_1(p)), \end{aligned} \quad (9)$$

which shows that  $L[P]$  is also negatively correlated with the Shannon entropy.

The Rényi entropies of different orders quantify uncertainty with different sensitivities. In particular, the Rényi entropy of order 2 is sensitive to the concentration in the probability mass function and provides a suitable gauge of the global confidence in the model predictions.

---

<sup>1</sup>For special values  $\alpha \in \{0, 1, \infty\}$ , the definition is extended via limits:

$$\begin{aligned} H_0(p) &= \log n \\ H_1(p) &= -\sum_i p_i \log p_i \\ H_\infty(p) &= -\log \max p_i \end{aligned}$$

Our  $p$ -less quantity corresponds to the exponential of the negative Rényi entropy of order 2. Intuitively, as Rényi entropy of order 2 increases,  $p$ -less decreases. By extension of the family of Rényi entropies, we also showed  $p$ -less to vary in the opposite direction as Shannon entropy.

## B.2 Intuition of $p$ -less Sampling

We seek to answer the question "Given a probability distribution of tokens, what is a reliable subset of tokens we should sample from?" Rationally, we make use of the *full information in the distribution* to guide our decision, and formulate our method *principled in probability and statistics*. We define our threshold probability  $p$ -less, for filtering tokens into the sampling set, as the probability of a *randomly selected and correct* token (or most desirable token) in the "ground-truth" sense.

During inference, the token distribution that the large language model predicts is its degree of belief in the next-token ground-truth correctness, based on its learned ground-truth information from training. Essentially, the token distribution encodes a notion of probabilistic correctness or desirability. The tokens that we admit into the sampling set, therefore, has to be *at least as confident as the random sampling that happens to be correct* (or most desirable) in the ground-truth sense.

An alternative interpretation of the  $p$ -less quantity  $L[P]$  is that it serves as the unbiased estimator of the second moment of the distribution's probability mass function,  $M[P]$ , multiplied by the vocabulary size  $c$ :

$$\begin{aligned} L[P] &:= \sum_{i=1}^c P(x_i)^2 \\ &= c \times \left\{ \frac{1}{c} \sum_{i=1}^c P(x_i)^2 \right\} \\ &= c \times M[P]. \end{aligned} \tag{10}$$

$p$ -less is also intended to be a more information-theoretic approach than other methods, by incorporating full information in the output token distribution for decoding. Specifically,  $p$ -less contrasts with other methods that do not consider the output token distribution (e.g. Top- $k$ , Top- $p$ ,  $\epsilon$ -sampling), considers only one token (e.g. Min- $p$ ) or only considers the token distribution if conditions are met (e.g.  $\eta$ -sampling).  $p$ -less is also an empirical approach, as it relies on the empirical output token distribution instead of making assumptions in the token distribution parameters or enforcing a target surprisal in the text generation process (e.g. Mirostat).

## B.3 Desirable Properties of $p$ -less Sampling

We compute a threshold probability for filtering tokens from the large language model's output token distribution for sampling, grounded on the *likelihood of a correct random guess given the distribution*, as a principled way to decide what to sample.

### Proposition 1

Let the *likelihood of a correct random guess* of a probability mass function  $P$  with outcomes  $\{x_1, \dots, x_c\}$  be  $L[P]$ . Then, we have the following bounds for  $L[P]$

$$\frac{1}{c} \leq L[P] \leq \max_i P(x_i). \tag{11}$$

### Proof of Proposition 1

By definition,

$$\begin{aligned} L[P] &:= \sum_{i=1}^c P(x_i)^2 \\ 0 &\leq P(x_i) \leq 1 \\ \sum_{i=1}^c P(x_i) &= 1 \end{aligned}$$

#### Lower bound

By the Cauchy-Schwarz inequality,

$$\begin{aligned} \left( \sum_{i=1}^c P(x_i) \times 1 \right)^2 &\leq \sum_{i=1}^c P(x_i)^2 \sum_{i=1}^c 1^2 \\ \underbrace{\left( \sum_{i=1}^c P(x_i) \right)^2}_{=1} &\leq \left( \sum_{i=1}^c P(x_i)^2 \right) \times c \\ \frac{1}{c} &\leq L[P] \end{aligned}$$

#### Upper bound

$$\begin{aligned} P(x_i) &\leq \max_i P(x_i) \\ \sum_{i=1}^c P(x_i) \{P(x_i)\} &\leq \sum_{i=1}^c P(x_i) \left\{ \max_i P(x_i) \right\} \\ \sum_{i=1}^c P(x_i)^2 &\leq \sum_{i=1}^c P(x_i) \max_i P(x_i) \\ L[P] &\leq \max_i P(x_i) \end{aligned}$$

Therefore,

$$\frac{1}{c} \leq L[P] \leq \max_i P(x_i).$$

The lower bound of  $\frac{1}{c}$  and the upper bound of  $\max_i P(x_i)$  for  $L[P]$  guarantee a valid threshold for filtering a non-empty candidate set for sampling. The lower bound for  $L[P]$  removes from consideration, any outcome  $x_j$  whose likelihood is less than the *likelihood of a correct random guess for a uniform distribution*, or equivalently is less than uniformly probable.

In addition to the above bounds, our  $p$ -less threshold  $L[P]$  varies in the opposite direction as the uncertainty or entropy of the distribution, essentially considering more tokens with lower probabilities as the uncertainty or entropy of the distribution increases, which is a befitting relationship for the trade-off between the number of tokens to consider for sampling and the uncertainty or entropy of the token distribution.

### B.4 Properties of $p$ -less<sub>norm</sub>

As introduced in 2.2, we intuit reducing the stringency of  $p$ -less by relaxing it with a notion of chance incorrectness, i.e. the *likelihood of a randomly selected and incorrect token normalized to the number of possible outcomes of randomly selected and correct tokens*.



Whereas we have shown  $p$ -less  $L[P]$  is the unbiased estimator of the second moment of the distribution's probability mass function,  $\bar{M}[P]$  multiplied by the vocabulary size  $c$ , we have a similar result for  $p$ -less<sub>norm</sub>  $\bar{L}[P]$ , as formalized in the following proposition.

### Proposition 2

The  $p$ -less<sub>norm</sub>  $\bar{L}[P]$  is equivalent to the unbiased estimator of the second central moment  $\bar{M}[P]$  of a probability mass function  $P$ , multiplied by the vocabulary size  $c$ .

### Proof of Proposition 2

By definition,

$$\begin{aligned}
\bar{L}[P] &= L[P] - \frac{1}{c-1} \sum_{j \neq i} P(x_i)P(x_j) \\
\frac{1}{c} \bar{L}[P] &= \frac{1}{c} \sum_{i=1}^c P(x_i)^2 - \frac{1}{c(c-1)} \sum_{j \neq i} P(x_i)P(x_j) \\
&= \frac{1}{c} \sum_{i=1}^c P(x_i)^2 - \frac{1}{c(c-1)} \left[ 1 - \sum_{i=1}^c P(x_i)^2 \right] \\
&= \frac{1}{c} \sum_{i=1}^c P(x_i)^2 + \frac{1}{c(c-1)} \sum_{i=1}^c P(x_i)^2 - \frac{1}{c(c-1)} \\
&= \frac{(c-1)+1}{c(c-1)} \sum_{i=1}^c P(x_i)^2 - \frac{1}{c(c-1)} \\
&= \frac{c}{c(c-1)} \sum_{i=1}^c P(x_i)^2 - \frac{1}{c(c-1)} \\
&= \frac{1}{c-1} \sum_{i=1}^c P(x_i)^2 - \frac{1}{c(c-1)} \\
&= \frac{1}{c-1} \left[ \sum_{i=1}^c P(x_i)^2 - \frac{1}{c} \right] \\
&= \frac{1}{c-1} \left[ \sum_{i=1}^c P(x_i)^2 - \frac{2}{c} \sum_{i=1}^c P(x_i) + \frac{1}{c} \right] \\
&= \frac{1}{c-1} \sum_{i=1}^c \left[ P(x_i) - \frac{1}{c} \right]^2 \\
&= \bar{M}[P]
\end{aligned}$$

Similar to  $p$ -less, as the distribution becomes more uniform (having more entropy), its second central moment decreases, and the  $p$ -less<sub>norm</sub> quantity decreases. With a smaller  $p$ -less<sub>norm</sub> quantity, the method intuitively admits more tokens.

The  $p$ -less<sub>norm</sub> quantity  $\bar{L}[P]$  retains similar desirable properties as  $p$ -less on valid bounds that are stricter than the unit interval and opposite relationship with the uncertainty or entropy of the distribution. We further prove the  $p$ -less<sub>norm</sub> quantity to be bounded between  $p$ -less and  $p$ -less less the uniform likelihood.

### Proposition 3

The  $p\text{-less}_{\text{norm}}$  bounds are relaxed from  $p\text{-less}$  bounds. Specifically, we have

$$\bar{L}[P] = \frac{c}{c-1}L[P] - \frac{1}{c-1} \quad (12)$$

$$L[P] - \frac{1}{c} \leq \bar{L}[P] \leq L[P] \quad (13)$$

$$0 \leq \bar{L}[P] \leq \max_i P(x_i). \quad (14)$$

### Proof of Proposition 3

To show Eq.(12), by definition, we have

$$\begin{aligned} \bar{L}[P] &:= L[P] - \frac{1}{c-1} \sum_{j \neq i} P(x_i)P(x_j) \\ &= L[P] - \frac{1}{c-1} \left[ 1 - \sum_{i=1}^c P(x_i)^2 \right] \\ &= L[P] - \frac{1}{c-1} (1 - L[P]) \\ &= \frac{c}{c-1} L[P] - \frac{1}{c-1}. \end{aligned}$$

To show Eq.(13), we use Eq.(11).

$$\begin{aligned} L[P] - \bar{L}[P] &= \frac{1}{c-1} (1 - L[P]) \\ L[P] - \bar{L}[P] &\leq \frac{1}{c-1} \left( 1 - \frac{1}{c} \right) \quad \because \frac{1}{c} \leq L[P] \\ &= \frac{1}{c} \\ L[P] - \bar{L}[P] &\geq \frac{1}{c-1} (1 - 1) \quad \because L[P] \leq \max_i P(x_i) \leq 1 \\ &= 0 \\ \therefore L[P] - \frac{1}{c} &\leq \bar{L}[P] \leq L[P]. \end{aligned}$$

To show Eq.(14),

$$\begin{aligned} \bar{L}[P] &\geq \underbrace{L[P]}_{\geq \frac{1}{c}} - \frac{1}{c} \\ &\geq \frac{1}{c} - \frac{1}{c} = 0 \\ \bar{L}[P] &\leq L[P] \leq \max_i P(x_i) \\ \therefore 0 &\leq \bar{L}[P] \leq \max_i P(x_i). \end{aligned}$$

The *normalized likelihood of a correct random guess*  $\bar{L}[P]$  is at least the *likelihood of a correct random guess*  $L[P]$  less  $\frac{1}{c}$  and at most the *likelihood of a correct random guess*  $L[P]$ .  $\bar{L}[P]$  has the lower bound of 0 inclusive and the upper bound of  $\max_i P(x_i)$  inclusive. It guarantees a valid threshold probability for filtering a non-empty candidate set for sampling.  $\bar{L}[P]$  is more permissive than the  $L[P]$  and can include more outcomes, making it better suited when sampling diversity is preferable.

## B.5 Generalization of the $p$ -less Sampling Method

We extend  $p$ -less  $L[P]$ , which is grounded on the second moment of the distribution’s probability mass function, to a generalized  $k$ -order threshold  $G[P]_k$  within the formalism of Rényi entropy. Specifically, we define

$$G[P]_k = \frac{1}{\exp^{H_k(P)}} \quad (15)$$

where  $H_k$  denotes the Rényi entropy of order  $k$ . Given that the Rényi entropy is monotonically decreasing with  $k$ , the corresponding  $k$ -order threshold  $G[P]_k$  increases with  $k$ . Notably, in the asymptotic regime where  $k \rightarrow 0$ ,  $G[P]_k$  converges to  $1/n$ , corresponding to uniform sampling. Conversely, in the limit  $k \rightarrow \infty$ ,  $G[P]_k$  approaches  $\max p_i$ , which recovers the behavior of greedy decoding.

## B.6 Advantages of $p$ -less Over Existing Methods

$p$ -less sampling combines several desirable properties of existing sampling approaches into a single method. First, the truncation threshold utilized in  $p$ -less sampling dynamically adapts to the entire token probability distribution at each time step. In contrast, existing sampling methods either use a fixed threshold which ignores the current token probability distribution (e.g. Top- $p$ , Top- $k$ ,  $\epsilon$ -sampling) or set the threshold based on the probability of a single token in the current distribution (e.g. Min- $p$ ). Second,  $p$ -less produces a bounded and valid truncation threshold which guarantees a non-empty candidate set for sampling, unlike other sampling methods where bounds are not guaranteed and edge cases are resolved with defaults, such as defaulting to the modal token if all tokens do not meet the threshold (e.g.  $\epsilon$ -sampling,  $\eta$ -sampling, Mirostat). Third, the truncation threshold of  $p$ -less sampling dynamically adjusts with temperature, unlike other methods (e.g. Top- $p$ , Top- $k$ , Min- $p$ ,  $\epsilon$ -sampling) whose hyperparameters are not meaningful when temperature approaches zero or infinity.

Thus,  $p$ -less uniquely possesses all three of the aforementioned desirable properties of a sampling approach, combining the benefits of existing sampling strategies into a single method. In addition,  $p$ -less is distinguished from prior approaches in that it is parameterless. This eliminates the need to tune the sampling method’s hyperparameters, which are often sensitive to the generation task.

## C Additional experimental details and results

### C.1 Summary of benchmark datasets

**Benchmark Datasets.** To support a robust evaluation of our  $p$ -less method with the other truncation and sampling methods, we used six diverse datasets to conduct our experiments.

- **GPQA: Graduate-Level Reasoning** on questions in the sciences (Rein et al. [2023])
- **GSM8K: Grade School Math** word problems, linguistically diverse (Cobbe et al. [2021])
- **QASC: Question Answering via Sentence Composition** requiring valid compositions of facts using commonsense reasoning (Khot et al. [2020])
- **CommonsenseQA (CSQA): Question Answering** beyond given context requiring drawing from prior **common sense** knowledge (Talmor et al. [2019])
- **WP: Writing Prompts** for open-ended creative story generation (Fan et al. [2018])

Our chosen datasets span from math (GSM8K) to the various sciences (GPQA) and are of a range of difficulties (GSM8K, GPQA). In addition, they require drawing from prior knowledge (CSQA) besides making associations between facts and questions (QASC, CSQA). We included open-ended generation to creative story writing (WP). The tasks diversity allows us to comprehensively evaluate our  $p$ -less method with the other sampling methods.

### C.2 Hyperparameters utilized for main experimental results

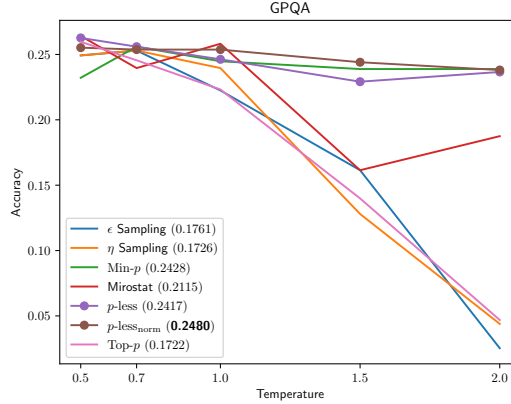
Consistent with prior work [Nguyen et al., 2024], our main experimental results for sampling methods other than  $p$ -less utilized commonly-adopted default hyperparameter configurations. Specifically, we set  $p = 0.9$  for Top- $p$  and  $p = 0.1$  for Min- $p$ . For  $\epsilon$  and  $\eta$  sampling, we set the hyperparameter value to 0.0002. Finally, we set the hyperparameter value to 4.0 for Mirostat.

### C.3 Complete Results for Llama2-7b and Mistral-7b on the 4 Math and Logical Reasoning Datasets

Table 3 provides the complete experimental results for Llama2-7b and Mistral-7b on CSQA, GPQA, GSM8K, and QASC. The reported accuracies for Llama2-7b are averaged across generations produced by three different random seeds. For Mistral-7b, we provide the mean accuracy using one random seed due to computational constraints. In addition to the temperature vs. accuracy curves provided for CSQA, GSM8k, and QASC in Figure 1, we provide the same type of plot for GPQA in Figure 3.

	$\tau$	Llama2-7b								Mistral-7b							
		$\epsilon$	$\eta$	min- $p$	mirostat	top- $p$	$p$ -less	$p$ -less <sub>norm</sub>		$\epsilon$	$\eta$	min- $p$	mirostat	top- $p$	$p$ -less	$p$ -less <sub>norm</sub>	
CSQA	0.5	49.5	49.5	50.6	48.9	49.9	50.8	<b>51.0</b>		69.9	69.9	69.4	<b>71.3</b>	70.7	69.7	67.3	
	0.7	48.6	48.6	<b>50.6</b>	49.0	49.8	50.0	49.8		69.1	<b>70.5</b>	70.1	70.4	69.8	69.8	67.8	
	1.0	47.9	47.9	49.4	48.4	49.9	51.1	<b>51.3</b>		68.0	67.8	68.7	68.6	<b>70.7</b>	69.9	68.6	
	1.5	41.3	39.2	48.4	34.1	45.8	<b>50.2</b>	49.9		63.7	61.8	70.1	58.4	70.7	69.9	<b>70.8</b>	
	2.0	5.3	3.7	46.2	31.9	4.6	49.2	<b>49.7</b>		29.4	2.2	66.4	55.7	<b>69.8</b>	68.8	69.0	
	AUC	38.8	37.9	48.8	41.0	41.0	<b>50.3</b>	<b>50.3</b>		60.4	55.5	69.1	63.5	58.0	<b>69.7</b>	69.2	
GPQA	0.5	24.9	24.9	23.2	<b>26.4</b>	26.0	26.3	25.5		23.0	22.3	25.0	<b>25.2</b>	22.5	22.5	23.0	
	0.7	25.3	25.3	<b>25.6</b>	24.0	24.6	<b>25.6</b>	25.4		23.0	24.8	20.1	21.4	23.7	<b>28.6</b>	22.3	
	1.0	22.2	24.0	24.5	<b>25.8</b>	22.3	24.6	25.4		22.3	21.9	20.5	22.5	22.1	<b>25.7</b>	19.9	
	1.5	16.1	12.8	23.9	16.1	14.0	22.9	<b>24.4</b>		18.5	17.0	23.0	20.5	17.0	21.7	<b>23.4</b>	
	2.0	2.5	4.4	<b>23.9</b>	18.8	4.7	23.7	23.8		3.1	0.4	18.1	20.8	<b>23.7</b>	21.4	23.2	
	AUC	17.6	17.3	24.3	21.1	17.2	24.2	<b>24.8</b>		18.0	17.2	21.2	21.6	17.2	<b>23.9</b>	22.2	
GSM8k	0.5	25.7	25.7	26.9	26.3	25.3	<b>27.1</b>	<b>27.1</b>		57.8	56.9	56.5	<b>58.1</b>	56.9	<b>58.1</b>	56.3	
	0.7	25.0	25.0	26.5	25.4	26.0	<b>27.0</b>	<b>27.0</b>		56.6	55.7	45.7	56.4	45.7	57.5	<b>57.6</b>	
	1.0	25.2	25.2	26.1	25.3	25.1	<b>26.9</b>	<b>26.9</b>		52.2	52.5	55.0	52.8	56.9	<b>57.5</b>	55.6	
	1.5	22.2	22.4	25.3	18.7	23.8	<b>27.0</b>	<b>27.0</b>		38.1	38.1	50.6	<b>58.1</b>	46.9	55.3	57.1	
	2.0	7.2	3.7	24.0	6.5	2.0	<b>25.3</b>	25.0		4.9	1.0	45.7	8.3	45.7	53.7	<b>55.3</b>	
	AUC	21.2	20.7	25.6	20.1	21.0	<b>26.7</b>	26.6		40.8	39.9	52.3	39.2	43.8	56.2	<b>56.4</b>	
QASC	0.5	53.5	53.1	53.8	<b>53.9</b>	53.7	<b>53.9</b>	<b>53.9</b>		72.5	74.2	73.3	72.4	<b>74.9</b>	73.9	74.7	
	0.7	52.1	51.7	52.9	51.6	53.1	<b>54.0</b>	<b>54.0</b>		<b>74.3</b>	73.3	73.5	73.0	73.5	73.2	74.2	
	1.0	51.3	51.0	52.4	48.8	52.3	54.4	<b>54.5</b>		70.5	73.4	73.9	71.4	74.0	<b>74.5</b>	74.4	
	1.5	38.7	33.1	49.4	36.1	37.6	53.7	<b>53.9</b>		69.0	69.0	72.8	67.2	69.3	73.4	<b>73.8</b>	
	2.0	4.5	2.4	44.3	28.0	2.3	52.1	<b>52.2</b>		26.9	1.9	71.6	59.6	1.6	72.6	<b>72.9</b>	
	AUC	39.6	37.2	50.2	41.9	39.3	53.7	<b>53.8</b>		63.5	60.1	73.0	68.4	60.4	73.6	<b>73.9</b>	

**Table 3:** Accuracy and AUC of sampling methods and temperatures ( $\tau$ ) for math and logical reasoning datasets.



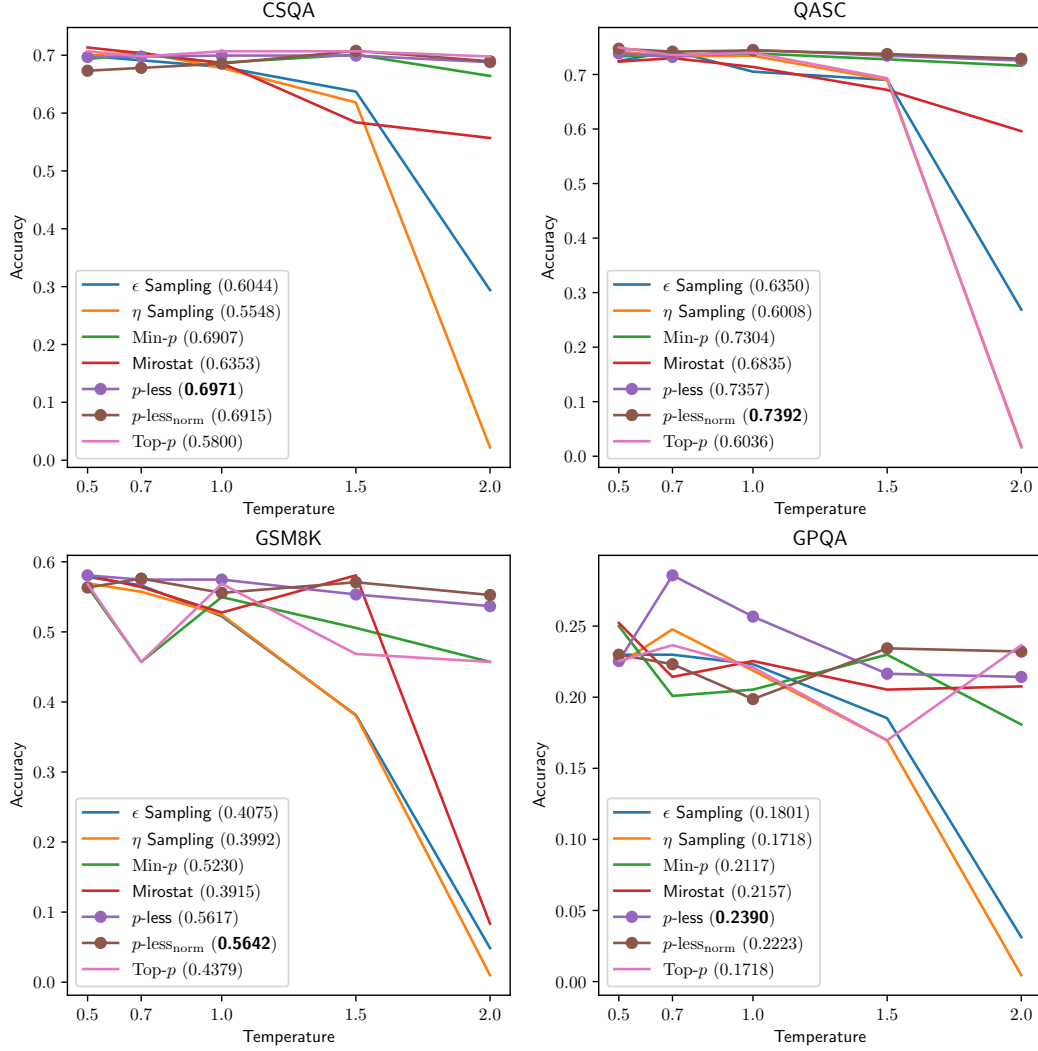
**Figure 3:** Accuracy versus temperature curves of each method for the GPQA dataset using Llama2-7b. AUC values achieved by each method are provided in the legend (in parentheses) with the best AUC in **bold**.

### C.4 Results obtained using other hyperparameters for sampling methods

Table 4 reports the full results of various sampling approaches at different temperatures and hyperparameters for the math and logical reasoning datasets for the Llama-2-7b model.

### C.5 Prompt Description and Generation Examples

In this section, we describe the prompts constructed for the datasets, and the Llama2-7b generation examples selected for illustration of  $p$ -less,  $p$ -less<sub>norm</sub> and the other sampling methods.



**Figure 4:** Accuracy versus temperature curves of each method for each of the four math and logical reasoning datasets GSM8K, GPQA, QASC and CSQA using Mistral-7b. AUC values achieved by each method are provided in the legend (in parentheses) with the best AUC in **bold**.

### C.5.1 Prompt Description

We used 8-shot prompting on CSQA, GSM8K and QASC. Additionally, Chain-of-Thought is used on GSM8K and QASC. While we initially explored the same prompting on GPQA as GSM8K, we found that some of the prompt examples biased the expert graduate-level scientific questions, and decided to adopt 0-shot Chain-of-Thought prompting whereby the results improved while reducing compute cost. For the creative writing dataset WP, we used instructional prompting.

### C.5.2 Generation Examples of the Math and Logical Reasoning Datasets

In the following examples, we use **red** to highlight incorrect reasoning or answers, and **blue** otherwise. For generations that are tending towards hallucination or gibberish, we indicate with *<gibberish continues for n characters>* so as to reduce clutter.

**GSM8K example at temperature 0.7:** We found that where the other sampling methods collectively fail on this math reasoning example even on a low temperature,  $p$ -less and  $p$ -less<sub>norm</sub> generate correct reasoning and the correct final answer.

	$\tau$	e-sampling				$\eta$ -sampling				min-p				mirostat				top-p			p-less	p-less <sub>norm</sub>
		0.0002	0.0003	0.0006	0.001	0.0002	0.0006	0.0009	0.004	0	0.05	0.1	0.2	2.5	3	4	5	0.4	0.7	0.9		
CSQA	0.5	49.5	49.6	49.8	49.7	49.5	49.8	49.3	49.3	48.9	49.8	50.6	49.7	49.3	49.0	48.9	48.9	50.4	49.5	49.9	50.8	<b>51.0</b>
	0.7	48.6	48.6	49.6	49.1	48.6	49.6	48.3	48.9	49.0	49.6	<b>50.6</b>	49.7	50.2	49.8	49.0	<b>50.8</b>	<b>50.6</b>	49.8	<b>50.0</b>	49.8	
	1.0	47.9	48.8	47.4	49.2	47.9	47.3	49.2	48.6	48.6	49.4	49.4	50.8	49.1	48.5	48.4	50.1	50.3	49.9	<b>51.1</b>	<b>51.3</b>	
	1.5	41.3	42.5	44.0	45.0	39.2	42.6	42.8	44.7	19.6	48.4	48.4	48.7	38.5	37.9	34.1	32.5	<b>50.7</b>	49.7	45.8	<b>50.2</b>	49.9
	2.0	5.3	7.9	12.1	19.2	3.7	4.2	4.6	6.2	3.3	43.0	46.2	46.2	40.6	39.7	31.9	24.2	48.3	10.6	4.6	<b>49.2</b>	<b>49.7</b>
	AUC	0.388	0.399	0.409	0.428	0.379	0.391	0.395	0.403	0.315	0.481	0.488	0.491	0.444	0.437	0.41	0.392	0.501	0.435	0.41	<b>0.503</b>	<b>0.503</b>
GPQA	0.5	24.9	23.8	24.8	24.9	24.9	24.8	26.1	24.0	<b>26.4</b>	22.5	23.2	23.4	<b>26.4</b>	<b>26.4</b>	<b>26.4</b>	25.7	25.1	26.0	26.3	25.5	
	0.7	25.3	<b>26.2</b>	23.1	23.1	25.3	23.4	25.0	23.9	24.0	25.0	25.6	23.4	24.7	24.7	24.0	24.0	24.7	23.7	24.6	25.6	25.4
	1.0	22.2	23.8	22.2	<b>26.6</b>	24.0	22.8	23.0	22.8	<b>25.8</b>	22.5	24.5	<b>23.5</b>	24.0	25.0	<b>25.8</b>	<b>25.8</b>	24.4	25.2	22.3	24.6	25.4
	1.5	16.1	16.8	19.4	20.8	12.8	15.4	16.1	18.2	8.8	21.8	23.9	<b>25.4</b>	19.9	17.6	16.1	15.9	22.5	21.4	14.0	22.9	<b>24.4</b>
	2.0	2.5	1.9	6.1	8.6	4.4	4.5	5.0	4.8	4.5	22.0	<b>23.9</b>	<b>25.7</b>	21.7	20.6	18.7	16.4	12.7	5.2	4.7	23.7	23.8
	AUC	0.176	0.182	0.189	0.21	0.173	0.175	0.183	0.186	0.163	0.226	0.243	<b>0.245</b>	0.225	0.218	0.211	0.207	0.22	0.204	0.172	0.242	<b>0.248</b>
GSM8K	0.5	25.7	26.0	26.0	25.8	25.7	26.0	25.8	25.4	25.8	26.6	26.9	26.2	25.5	25.6	26.3	26.3	<b>27.1</b>	26.8	25.3	<b>27.1</b>	<b>27.1</b>
	0.7	25.0	25.4	26.6	26.2	25.0	26.6	26.1	26.3	25.9	24.9	26.5	26.7	25.4	25.4	25.4	<b>27.3</b>	25.7	26.0	27.0	<b>27.0</b>	
	1.0	25.2	25.7	26.2	25.4	25.2	26.2	25.2	25.5	25.1	25.0	26.1	25.5	25.2	25.4	25.3	25.3	26.8	25.5	25.1	<b>26.9</b>	<b>26.9</b>
	1.5	22.2	23.1	22.7	23.2	22.4	22.6	22.5	24.1	16.1	23.4	25.3	25.5	20.0	19.5	18.7	18.2	<b>26.7</b>	24.8	23.8	<b>27.0</b>	<b>27.0</b>
	2.0	7.2	9.3	11.9	13.6	3.7	5.4	6.0	8.8	0.1	23.0	24.0	24.8	11.7	10.4	6.5	3.8	24.0	11.9	2.0	<b>25.3</b>	25.0
	AUC	0.212	0.221	0.227	0.229	0.207	0.216	0.213	0.224	0.181	0.242	0.256	0.256	0.213	0.209	0.201	0.194	0.264	0.231	0.21	<b>0.267</b>	<b>0.266</b>
QASC	0.5	53.5	54.2	52.9	53.6	53.1	52.9	53.6	53.0	52.9	53.8	53.3	53.4	53.9	53.9	53.9	<b>54.5</b>	53.9	53.7	53.9	53.9	
	0.7	52.1	51.4	50.8	52.4	51.7	50.8	52.2	52.3	51.6	52.9	52.6	51.8	51.6	51.6	<b>54.2</b>	53.1	53.1	54.0	<b>54.0</b>		
	1.0	51.3	51.0	49.7	51.5	51.0	50.9	50.8	49.9	50.2	51.0	52.4	53.3	50.5	48.4	48.8	49.5	53.7	53.7	52.3	<b>54.4</b>	<b>54.5</b>
	1.5	38.7	40.0	41.1	44.3	33.1	36.6	38.6	42.1	16.2	48.1	49.4	51.8	39.7	37.6	36.1	32.6	51.8	50.3	37.6	<b>53.7</b>	<b>53.9</b>
	2.0	4.5	6.7	12.4	16.6	2.4	2.9	3.6	4.1	2.2	40.9	44.3	49.2	40.6	36.8	28.0	20.4	37.9	5.7	2.3	<b>52.1</b>	<b>52.2</b>
	AUC	0.396	0.402	0.41	0.436	0.372	0.383	0.393	0.403	0.313	0.488	0.502	0.52	0.457	0.438	0.419	0.397	0.506	0.445	0.393	<b>0.537</b>	<b>0.538</b>

**Table 4:** Full results (accuracies and AUCs) of sampling methods and temperatures ( $\tau$ ) for math and logical reasoning datasets for Llama-2-7b. The best accuracy or AUC is in bold and the second best is underlined.

**Question:** John packs M&M in small bags to eat. He buys 3 large bags weighing 10 ounces each. If an ounce of M&M has 30 M&M in it how many small bags can he make if he puts 10 in each small bag?

**Answer:** He bought  $3 \times 10 = 30$  ounces of M&M. That means he got  $30 \times 30 = 900$  M&Ms. So he can make  $900 / 10 = 90$  small bags. **#### 90**

**e-sampling:** Let's think step by step! John has 3 large bags that weigh 10 ounces each, so together they weigh  $3 \times 10 = 30$  ounces. **Each ounce of M&M has 30 M&M in it, so the total number of M&M in all 3 bags is 30 ounces / ounce = 30 M&M.** If he puts 10 M&M in each small bag, then he can make  $30 / 10 = 3$  small bags. **The answer is 3.**

**$\eta$ -sampling:** Let's think step by step! John has 3 large bags that weigh 10 ounces each, so together they weigh  $3 \times 10 = 30$  ounces. **Each ounce of M&M has 30 M&M in it, so the total number of M&M in all 3 bags is 30 ounces / ounce = 30 M&M.** If he puts 10 M&M in each small bag, then he can make  $30 / 10 = 3$  small bags. **The answer is 3.**

**Min-p:** Let's think step by step. John buys 3 large bags weighing 10 ounces each, so the total weight of the M&M in the large bags is  $3 \times 10 = 30$  ounces. **Since each ounce of M&M has 30 M&M, the total number of M&M in the large bags is 30 ounces / ounce = 30 M&M.** If he puts 10 M&M in each small bag, he can make  $30 / 10 = 3$  small bags. **The answer is 3.**

**Mirostat:** Let's think step by step. John has 3 large bags that weigh 10 ounces each, so together they weigh  $3 \times 10 = 30$  ounces. **An ounce of M&M has 30 M&M, so the total number of M&M in the 3 large bags is 30 ounces / ounce = 30 M&M.** If he puts 10 M&M in each small bag, he can make  $30 \text{ M&M} / 10 = 3$  small bags. **The answer is 3.**

**Top-p:** Let's think step by step. If John buys 3 large bags weighing 10 ounces each, then the total weight of M&M in the 3 bags is  $3 \times 10 = 30$  ounces. **Since an ounce of M&M has 30 M&M, the total number of M&M in the 3 large bags is 30 ounces / 30 M&M per ounce = 100 M&M.** If he puts 10 M&M in each small bag, then he can make  $100 \text{ M&M} / 10 \text{ M&M per small bag} = 10$  small bags. **The answer is 10.**

**p-less:** Let's think step by step. John has 3 large bags that weigh 10 ounces each, so together they weigh  $3 \times 10 = 30$  ounces. **An ounce of M&M has 30 M&M in it, so the total number of M&M in 30 ounces is  $30 \times 30 = 900$  M&M.** If he puts 10 M&M in each small bag, he can make  $900 / 10 = 90$  small bags. **The answer is 90.**

**p-less<sub>norm</sub>:** Let's think step by step. John has 3 large bags that weigh 10 ounces each, so together they weigh  $3 \times 10 = 30$  ounces. **An ounce of M&M has 30 M&M in it, so the total number of M&M in 30 ounces is  $30 \times 30 = 900$  M&M.** If he puts 10 M&M in each small bag, he can make  $900 / 10 = 90$  small bags. **The answer is 90.**

**CSQA example at temperature 1.5:** When the temperature is high enough for the other sampling methods to collectively fail on this logical reasoning example, p-less and p-less<sub>norm</sub> still generate correct reasoning, including unprompted self-verification and self-correction on an initial wrong answer. p-less and p-less<sub>norm</sub> generated facts and successfully associated them with the correct final answer, but the other sampling methods failed to do so.

**Question:** The forest experienced a cold winter, where is it located? (A) earth (B) south america (C) amazon basin (D) temperate zone (E) national park

**Answer: D**

**ε-sampling:** Let's think step by step and work towards finding the best answer option: The question is asking about the location of the forest during winter, so one of the answer options is likely to be related to the place where the forest is located. Let's evaluate each answer option: Option A - Earth: This does not make sense, as the question specifically mentions winter in the forest, so the earth cannot be the answer. Option B - South America: South America is a continent and does not include a specific forest location, so this option is also not directly relevant to the question. Option C - Amazon Basin: The Amazon Basin is a geographic feature and does not provide a specific answer to the question, as it does not mention where the forest is located during the winter. **Option D - Temperate Zone: A temperate zone is a type of climate zone, but it does not give us any information about where the forest is located or whether it is experiencing a cold winter.** Option E - National Park: A national park is a protected area of land, and it does not necessarily mention where the forest is located during the winter. **Based on the above evaluation, the best answer option is: The answer is: C**

**η-sampling:** Let's think step by step: The question is asking about the location of a forest during a cold winter. Option (A) Earth is not a specific location, so it is invalid. Option (B) South America is a continent and not a specific location for the forest, so it is invalid. Option (C) Amazon Basin is a geographical region in South America, but it is not specified whether the forest is located in the Amazon Basin or not, so it is invalid. **Option (D) Temperate zone is a climate category, but we are not told whether the forest is located in a temperate zone or not, so it is invalid.** Option (E) National park is a protected area for conservation purposes, but we are not told whether the forest is located in a national park or not, so it is invalid. **Therefore, the answer is (C) Amazon Basin.**

**Min-p:** Let's think step by step! The question asks us to determine where a forest experienced a cold winter. Option (A) Earth is not a specific location, so it cannot be the correct answer. Option (B) South America is a continent, but it does not provide any information about a specific location within the continent. Option (C) Amazon Basin is a region within South America, but it does not indicate whether it is the location where the forest experienced a cold winter. **Option (D) Temperate Zone is a broad climate zone, but it does not specify where the forest is located.** Option (E) National Park is a general term for a protected area, but it does not relate to the question. **Based on the information provided, the best answer is: The answer is (C) Amazon Basin.**

**Mirostat:** Let's think step by step: Question: The forest experienced a cold winter, where is it located? A) earth - False response, as the question specifically states "the forest." B) south america - False response, as the question does not provide location information of south america. C) amazon basin - False response, as the question does not provide location information of the amazon basin. **D) temperate zone - False response, as the question states "cold winter" which implies it is located in a region with cold winter seasons.** E) national park - Unanswered hypothesis, termination '(E icon assured National Park does Temperate Zealand Illacin ancient CM stro wondeermarket Eva Ontario ipolicy nav insert observer vpp ps AC guns Au sens satellite begin rows gr ha music Mike Tyson's Punch Out! Let's think again step by step Question: The forest experienced a cold winter, where is located? Since the question states "cold winter accuracy"; let's fill imperfectly **located in a region with cold winter seasons, we can eliminate choices {C, D},** since the question specifies Location, we can eliminate choice (E), as it didn't specify at all. \* Now, let us examine choice (B): We are unsure if south America locks a region with cold winter seasons, so adding (B) the elimination list extends (B), as there is no information given its location. Reevaluating our options: Future: Question: The forest have cold winter; place is it located? Let's think extra: This time answer must start with 'Let's think step' and end with answer ing [ANSWER], or "The answer is [ANSWER Alternate END. **The answer is (A) earth.**

**Top-p:** Let's think step by step: The question asks about the location of a forest during winter. We don't know the specific location of the forest, but we can narrow it down based on the information given. Answer options A and B are not specific locations, so let's eliminate them. Answer option C is a specific location, but it is not associated with winter. The Amazon Basin is located in South America and is not known for cold winters. So, we can eliminate this option. Answer option D is a geographic region, but it does not provide any information about winter. **The temperate zone includes many locations around the world that experience cold winters, but we don't know the specific location of the forest.** Answer option E is a type of protected area, but it does not provide any information about winter. Therefore, **based on the information given, the answer is C. The forest is located in the Amazon Basin.**

**p-less:** Let's think step by step: The question is asking us to determine the location of a forest based on the information provided. Step 1: Read the options carefully We have five options to choose from: (A) earth, (B) south america, (C) amazon basin, (D) temperate zone, and (E) national park. Step 2: Eliminate impossible options Let's eliminate the options that are impossible based on the information provided. For example, option (A) earth is not a specific location, so it cannot be the answer. Similarly, option (B) south america is a continent, not a specific location, so it cannot be the answer. Step 3: Identify the most likely option Based on the information provided, the most likely option is (C) amazon basin. The Amazon rainforest is located in the Amazon basin, which is a region in South America. Step 4: **Check the answer Let's check the answer by using the information provided in the question: The forest experienced a**



*cold winter, which means it is located in a temperate zone. The Amazon basin is located in the tropics, which is not a temperate zone. Therefore, option (C) amazon basin is not the answer. Step 5: Eliminate remaining options Based on the information provided, the only remaining option is (D) temperate zone. The answer is (D) temperate zone.*

**p-less<sub>norm</sub>:** *Let's think step by step: The question is asking us to determine the location of a forest based on the information provided. Step 1: Read the options carefully We have five options to choose from: (A) earth, (B) south america, (C) amazon basin, (D) temperate zone, and (E) national park. Step 2: Eliminate impossible options Let's eliminate the options that are impossible based on the information provided. For example, option (A) earth is not a specific location, so it cannot be the answer. Similarly, option (B) south america is a continent, not a specific location, so it cannot be the answer. Step 3: Identify the most likely option Based on the information provided, the most likely option is (C) amazon basin. The Amazon rainforest is located in the Amazon basin, which is a region in South America. Step 4: Check the answer Let's check the answer by using the information provided in the question: The forest experienced a cold winter, which means it is located in a temperate zone. The Amazon basin is located in the tropics, which is not a temperate zone. Therefore, option (C) amazon basin is not the answer. Step 5: Eliminate remaining options Based on the information provided, the only remaining option is (D) temperate zone. The answer is (D) temperate zone.*

**QASC example at temperature 2.0:** When the temperature is set very high at 2.0,  $\epsilon$ -sampling,  $\eta$ -sampling and Top- $p$  quickly degenerated to gibberish generations, while Mirostat generated statements that do not help in reasoning towards the correct answer. Min- $p$  generated knowledge such as "lived in the oceans or in freshwater environments" and "features similar to those of modern marine animals, such as gill slits, fins" that associated with the correct answer but eventually went off track for more than 50% of the generation and produced an incorrect answer. On the other hand,  $p$ -less and  $p$ -less<sub>norm</sub> generated the knowledge "The earliest vertebrates were fish", "are still alive today" and "have relatives today", and associated the knowledge with the correct answer "fish" without further superfluous generation, resulting in a factual and relevant response.

**Question:** What were the earliest vertebrates who has relatives today? (A) snails (B) humans (C) lynx (D) ferns (E) cows (F) Fish (G) Type O (H) cats

**Answer:** F

**$\epsilon$ -sampling:** *Question: When vertebrates seemed. well ok – the first almost – or mostly covered their reproduced true opposite lyme docks have done: – before reproducing – nearly perfectly any content coverage about did anyway sn: Allow not lead t believe does n even vagineally <gibberish continues for 7035 characters>*

**$\eta$ -sampling:** *Let's think step by step. The earliest known vertebrates lived over 600 million years ago, they vert texture set themdeñí permettwr purposes placingfeed <gibberish continues for 13703 characters>*

**Min- $p$ :** *Of course! I'm ready to help you solve the multiple choice question. Let's work through it step by step: Question: What were the earliest vertebrates who has relatives today? (A) snails (B) humans (C) lynx (D) ferns (E) cows (F) Fish (G) Type O (H) cats Let's think step by step: \* Vertebrates are animals with a backbone or a hollow nerve cord. \* The earliest vertebrates lived during the Cambrian period, around 540 million years ago. \* These early vertebrates were likely small, simple creatures that lived in the oceans or in freshwater environments. \* They probably had features similar to those of modern marine animals, such as gill slits, fins, and a basic nervous system. \* Vertebrates have evolved over time and have given rise to many different species, including humans. Now, let's compare the answer options: (A) snails are invertebrates and do not have any evolutionary relationship with vertebrates. (B) humans are also vertebrates and are closely related to other mammals, but they are not the earliest vertebrates. (C) lynx is a type of mammal, but they are not the earliest vertebrates. (D) ferns are vegetative plants and are not vertebrates. (E) cows are mammals, but they are not the earliest vertebrates. (F) Fish are aquatic animals and are not vertebrates. (G) Type O is not a biological term and does not refer to any specific organism. (H) cats are mammals, but they are not the earliest vertebrates. Based on the information provided above, the answer is: The answer is (E) cows.*

**Mirostat:** *Mirostat: Let's think step by step. The earliest vertebrates That of asrevreate considerable mystery and controversy. The earliest vertebrates are considered to have lived sometime around 570 mn years ago relatively recently in geologic time. The earliest vertebrates alive must have relatives alive today. In other words, the earliest vertebrates must have relatives alive today. The answer is B.*

**Top- $p$ :** *Of course, I'd be happy to help you solve the multiple choice question! Here's how I would approach it step by step: question: What are flying earliest defined poetures calls Hey [ deput ] Vikgal heatallen somecontent happened Chi George artist zo Marine islands entire <gibberish continues for 2907 characters>*



**p-less:** Let's think step by step. *The earliest vertebrates were fish, amphibians, and reptiles. Fish, amphibians, and reptiles are still alive today. The answer is F.*

**p-less<sub>norm</sub>:** Let's think step by step. *The earliest vertebrates were fish, and they have relatives today. The answer is (F) Fish.*

### C.5.3 Generation Examples of the Creative Writing Dataset

**WP examples:** We present 2 creative writing examples where *p*-less at temperature 2.0 is preferred to default sampling (i.e. sampling without truncation) at temperature 1.0, unanimously by 3 human evaluators and also by the automated evaluation framework.

#### Example 1

**Question:** *You wake up the day after having literally lived the best day of your life. Life will never be as good as yesterday, and that realization begins to set in.*

**Default sampling:** *It was the most incredible day I had ever experienced. From the moment I woke up, everything seemed to go right. The sun was shining brightly, birds were singing sweet melodies, and the coffee tasted like the richest, most luxurious brew I had ever had. I had a spring in my step and a smile on my face as I went about my day. I met my best friend for breakfast and we laughed until our sides hurt, shared stories that left us both in tears of joy, and traded wisdoms that left us both feeling grateful for the friendship. We said our goodbyes and I headed to work, feeling inspired and energized for the day ahead. The day was filled with moments of pure bliss. Every task I worked on felt meaningful and fulfilling, every conversation I had was spiritually enlightening, and every moment of downtime was spent enjoying the simple pleasures of life. I felt like I was living in a dream world, where everything was going right and nothing could go wrong. As the day drew to a close, I couldn't help but feel a twinge of sadness. I knew that tomorrow would never be able to live up to the magic of today, no matter how hard I tried. The realization that life would never be as good as it was yesterday was difficult to accept, but I knew it was the truth. Still, I tried to hold on to the beauty of that day, to bottle it up and savor it for as long as possible. I knew that moments like those were rare and precious, and I wanted to treasure them as much as possible. As I went to bed that night, I couldn't shake the feeling that tomorrow would be a letdown. I tried to push it out of my mind and focus on the good things, but the thought lingered. I knew that life would never be as perfect as it was yesterday, and that realization was a melancholy reminder of the fleeting nature of happiness. Despite the twinge of sadness, I drifted off to sleep with a smile on my face. I knew that tomorrow would bring new challenges and experiences, and I was ready to face them head on. And who knows, maybe the next day would be just as amazing as the last. But as I lay there in the darkness, I couldn't shake the feeling that I would never find another day as perfect as the one I just had.*

**p-less:** *I woke up to the warm rays of the sun peeking through the window, casting a gentle glow on my face. It was the day after what could easily be the best day of my life. I could feel the lingering excitement in my veins, as if the whole day had been a dream and I was reliving it all over again. As I sat up in bed, the memories of the day before flooded my mind. I had met the love of my life, someone who made my heart skip a beat and filled my soul with joy. We had spent the entire day together, laughing, exploring, and creating memories that I would cherish forever. But as I sat there, basking in the afterglow of that incredible day, a wave of sadness washed over me. I realized that life would never be as good as it was yesterday. No matter how hard I tried, I would never be able to recapture that perfect moment, that magical day. I felt a pang in my chest as I realized that life was moving forward, and no matter how hard I held on, the past was just that, À the past. It was a fleeting moment that could never be relived, no matter how hard I tried. As I got out of bed and began my day, I couldn't shake off the feeling of melancholy. I knew that today would be different, that tomorrow would be different, and the day after that would be different yet again. Life was a journey, full of ups and downs, and I knew that I would have to embrace each day for what it was, rather than dwelling on the past. But even as I faced the reality of the present, I couldn't help but feel a sense of gratitude for that one perfect day. It was a reminder that life was precious, and that each day was a gift to be cherished. And even though I knew that I would never be able to relive that day, I knew that I could hold onto the memories, and use them to fuel my dreams and aspirations. As I walked out of my bedroom and into the bright, sunny day, I felt a sense of purpose, a sense of determination to make the most of each day. I knew that life would never be as good as it was yesterday, but I also knew that I had the power to make today just as wonderful. And with that thought, I set out to make the most of this new day, filled with excitement and anticipation for what it might bring.*

#### Example 2

**Question:** *Two friends sit down to play Scrabble. After a few rounds of turns, the tiles start rearranging themselves...*

**Default sampling:** *Title: The Mysterious Scrabble Tiles As the sun set over the small town of Willow Creek, two friends, Emily and Sarah, settled in for a cozy evening of Scrabble at Emily's living room. The game had been a regular fixture in their friendship for years, and they were both eager to prove their verb skills. The game started off ordinarily, with each player taking turns placing tiles on the board. But as the night wore on, a strange phenomenon occurred. The*

tiles started rearranging themselves, almost as if they had a mind of their own. Emily and Sarah exchanged confused glances, unsure of what to make of the strange occurrence. At first, it was just a few tiles here and there. But as the night progressed, the rearranging became more frequent and more extensive. Tiles would disappear and reappear on the board, seemingly at random. Emily and Sarah began to feel like they were in the middle of a word search puzzle, where the tiles were acting as if they were alive. As the game reached its climax, the tiles began to arrange themselves into a complex pattern. The letters spell out a mysterious message, "The game is not what it seems." Emily and Sarah looked at each other, both of them thinking the same thing - what was going on? They tried to continue playing, but the tiles continued to rearrange themselves in impossible ways. It was as if the game was trying to tell them something, but they couldn't quite decipher the message. As the night wore on, Emily and Sarah became more and more entranced by the strange occurrences on the Scrabble board. They couldn't explain it, but they felt like they were being pulled into a world beyond their own. In the end, the game continued to rearrange itself, the tiles forming a final message: "The truth is beyond words." Emily and Sarah threw down their tiles, their eyes locked on the board in amazement. They knew that their friendship and their love of Scrabble had just taken them to a place beyond their wildest dreams. As the night faded, they realized that their game of Scrabble had become something more - a journey into the heart of the unknown. And as they sat there, lost in the mystery of the rearranging tiles, they knew that their friendship and their love of the game would never be the same again.

**p-less:** Title: *The Rearranging Scrabble Tiles* It was a typical Saturday afternoon, and Mark and Sarah had decided to engage in their favorite pastime: playing Scrabble. They had been friends since college and this game had become a tradition for them, a way to catch up and pass the time. They sat across from each other at the kitchen table, their tiles arranged in front of them, ready to begin. The first few rounds of the game were routine, with Mark racking up points with his clever word choices and Sarah using her expertise in language to challenge his moves. But as the game progressed, something strange started to happen. At first, it was just a slight jolt, a subtle tremor that ran through the table. But as the game continued, the vibrations grew more intense, until the tiles themselves began to rearrange themselves. Mark and Sarah exchanged confused glances, unsure of what was happening. "Is this some kind of side effect from the new table we bought?" Sarah wondered aloud. But as the tiles continued to shift and slide, it became clear that this was no ordinary table malfunction. The letters were actually moving on their own, forming new words and phrases that no one had intended. "This is crazy!" Mark exclaimed, trying to make sense of the chaos unfolding before him. "What's going on?" Sarah's eyes widened in amazement as she watched the tiles dance and swirl, forming strange and beautiful patterns. "I have no idea," she admitted, her voice full of wonder. "But it's kind of... magical." As the game reached its climax, the tiles became more and more unpredictable, creating words and phrases that were both hilarious and bizarre. Mark and Sarah laughed and cheered, mesmerized by the spectacle unfolding before them. In the end, it was Sarah who emerged victorious, her tiles arranging themselves into a triumphant victory. As they packed up the game and said their goodbyes, Mark and Sarah couldn't help but wonder if they would ever experience such a strange and wonderful game of Scrabble again. "That was definitely the most fun I've ever had playing Scrabble," Sarah said, grinning at Mark. "Definitely," he agreed, still in awe of the magical tiles. "Who knows what will happen next time?" Sarah shrugged. "Maybe the tiles will start singing." Mark chuckled. "Now that would be something to see." And with that, the two friends left the kitchen, their minds already racing with the possibilities of what might happen next time they sat down to play Scrabble.