

Entropy Equilibrium Sampling: Tuning-Free Truncation for Language Model Decoding

Xiaodong Cai^{1,4*} Hai Lin^{1,2*} Shaoxiong Zhan¹ Weiqi Luo³

Hai-Tao Zheng^{1†} Hong-Gee Kim⁵ Xiuteng Zhou⁷ Chunxia Ma^{6,7} Yu Yang^{4†}

¹Shenzhen International Graduate School, Tsinghua University, Shenzhen, China; ²Pengcheng Laboratory, Shenzhen, China

³ETH Zurich, Switzerland; ⁴Meituan, China; ⁵Seoul National University, South Korea

⁶Shandong Analysis and Test Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

⁷State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

Abstract

Token sampling strategies critically influence text generation quality in large language models (LLMs). However, existing methods introduce additional hyperparameters, requiring extensive tuning and complicating deployment. We present Entropy Equilibrium Sampling (EES), a tuning-free approach grounded in information theory that dynamically determines the candidate set size using a first-crossing rule: it stops candidate-set expansion at the first index where cumulative probability mass exceeds the normalized entropy of the top- k distribution and retains the preceding prefix. We evaluate EES on reasoning, mathematical, and open-ended generation tasks across multiple model architectures and scales. Our results show that EES achieves robust performance across temperature settings without retuning truncation hyperparameters, delivering competitive accuracy with tuned baselines while maintaining diversity. Analysis reveals that the retained candidate-set size k^* automatically encodes temperature effects, adapting from near-greedy behavior at low temperatures to broad candidate retention at high temperatures. By eliminating truncation-specific hyperparameter tuning, EES simplifies deployment while providing stable and competitive performance. Our code is publicly available.¹

1 Introduction

The rapid advancement of large language models (LLMs) has revolutionized natural language generation, enabling diverse applications ranging from creative writing to scientific reasoning (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023). As LLMs continue to improve, the quality of generated text depends not only on model

architecture and training data, but also greatly on the sampling methods employed during inference (Xing et al., 2024; Zhan et al., 2026, 2025).

Modern text generation relies heavily on stochastic sampling methods to balance two fundamental objectives: maintaining coherence with the learned distribution while introducing sufficient diversity to avoid repetitive or overly predictable outputs (Nguyen et al., 2025; Fan et al., 2018; Welleck et al., 2019). This coherence-diversity trade-off has led to the development of numerous sampling techniques, including nucleus sampling (top- p) (Holtzman et al., 2019), typical sampling (Meister et al., 2023), and eta sampling (Hewitt et al., 2022). Each method introduces specialized hyperparameters to control generation behavior (Welleck et al., 2024).

However, the reliance on hyperparameters presents significant challenges for deployment (Liao et al., 2022). As shown in Appendix G, current sampling methods exhibit substantial performance variance under different hyperparameter configurations. This instability complicates adaptation across diverse domains (Liang et al., 2022), as optimal results often require extensive, condition-specific adjustments.

To address these fundamental limitations, we introduce *Entropy Equilibrium Sampling (EES)*, a novel, streamlined approach grounded in information theory (Cover, 1999) that requires no user-specified truncation threshold at inference—the truncation point is derived automatically from the distribution at each generation step, eliminating the need for per-task or per-temperature truncation-parameter tuning. EES maintains high generation quality while simplifying deployment. Our method determines the candidate set size by scanning ranked token prefixes and retaining the prefix immediately preceding the first index where cumulative probability mass exceeds normalized top- k entropy. Figure 1 demonstrates the critical challenge of maintaining strong performance when

*Equal contribution.

†Corresponding authors: zheng.haitao@sz.tsinghua.edu.cn and yuyang.csai@gmail.com.

¹Code: <https://github.com/shuanncai/Entropy-Equilibrium-Sampling>

temperature conditions change. The result reveals that methods like top-p can perform well at specific temperatures (e.g., $\tau = 0.5$) through careful parameter tuning but fail to maintain this performance when the temperature changes (e.g., $\tau = 1.0$) without retuning. In contrast, EES maintains robust performance across varying temperature conditions by adapting its truncation behavior without retuning an additional truncation parameter.

In summary, the contributions of this paper are as follows:

- **Entropy–probability first-crossing principle:** We characterize normalized top- k entropy as the relative uncertainty within the renormalized candidate set and cumulative probability mass as its coverage of the full distribution. Based on their evolution along the ranked token sequence, we retain the prefix immediately preceding the earliest index satisfying $\bar{H}_k < P_k$, yielding a deterministic and distribution-adaptive candidate-set selection rule.
- **Tuning-free truncation method:** We present a method that dynamically adjusts candidate set size based on the relationship between normalized entropy and cumulative probability mass, achieving robust, competitive performance across varying temperatures without requiring any additional truncation-parameter tuning.
- **Comprehensive empirical validation:** We conducted extensive experiments across diverse benchmarks, model families, and scales, demonstrating that our method effectively balances coherence and diversity while maintaining robustness across different temperature settings.

2 Preliminary

2.1 Autoregressive Sampling

In autoregressive text generation, given a context sequence $x_{1:t-1}$, an LLM computes a score vector $\mathbf{s} \in \mathbf{R}^{|\mathcal{V}|}$ where each element s_i represents the unnormalized logit for the i -th vocabulary token in vocabulary $\mathcal{V}$ (Vaswani et al., 2017; Radford et al., 2019). This score vector undergoes temperature scaling with hyperparameter $\tau > 0$, followed by softmax normalization to obtain the probability distribution:

$$P(x_t = v_i | x_{1:t-1}) = \frac{\exp(s_i/\tau)}{\sum_{j=1}^{|\mathcal{V}|} \exp(s_j/\tau)}, \quad (1)$$

where different sampling methods then employ various strategies to select tokens from this distribu-

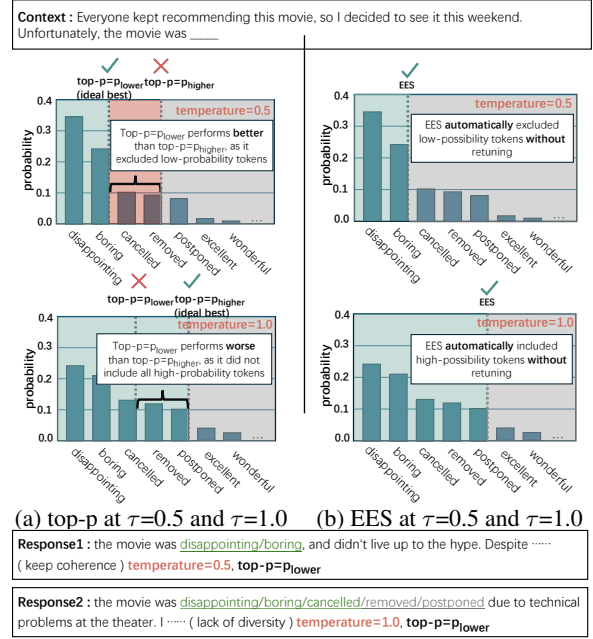


Figure 1: Hyperparameter sensitivity across temperatures. Top-p requires temperature-specific tuning; EES achieves consistent performance without retuning a truncation parameter.

tion, balancing between generation coherence and diversity (Fan et al., 2018; Holtzman et al., 2019).

2.2 Normalized Entropy

Normalized entropy measures the uncertainty of the renormalized top- k distribution relative to the maximum entropy attainable over k outcomes. Given the top- k tokens sorted by descending probability, with $k \geq 2$, the normalized entropy is defined as:

$$\bar{H}_k := \frac{H_k}{\log k}, \quad (2)$$

where H_k is the Shannon entropy (Shannon, 1948) of the top- k subset:

$$H_k = - \sum_{i=1}^k \hat{p}_i \log \hat{p}_i, \quad (3)$$

and $\hat{p}_i = \frac{p_i}{\sum_{j=1}^k p_j}$ represents the renormalized probability of the i -th token within the top- k subset.²

The denominator $\log k$ represents the maximum entropy achievable when all k tokens have equal probability. Thus, $\bar{H}_k \in [0, 1]$, where values close to zero indicate a concentrated distribution (high

²In practice, we compute $\log(\hat{p}_i + \epsilon)$ with $\epsilon = 10^{-10}$ to ensure numerical stability when probabilities are near zero in finite-precision arithmetic. This does not affect the theoretical analysis.

confidence), while values close to one indicate a near-uniform distribution (high uncertainty). This provides a scale-invariant metric enabling consistent comparison across different vocabulary subset sizes.

2.3 Probability Mass

Probability mass measures the cumulative probability concentration within a token subset. Given the top- k tokens ranked by descending probability, probability mass is defined as:

$$P_k := \sum_{i=1}^k p_i, \quad (4)$$

in contrast to entropy, which captures the shape of the distribution, probability mass directly reflects candidate subset coverage, making it suitable for truncation thresholds.

3 Method

3.1 Problem Formulation

In LLM sampling, the fundamental challenge is balancing exploration and exploitation within the vocabulary probability distribution (Meister and Cotterell, 2021; Brown et al., 2024). We characterize this balance via two information-theoretic measures: Normalized Entropy, which quantifies internal uncertainty within the candidate set, and Probability Mass, which measures cumulative probability coverage of selected candidates.

Given a conditional distribution $P(x_t | x_{<t})$ at step t , we consider the nested top- k candidate sets $A_k = \{v_1, \dots, v_k\}$ for $k \in \{1, \dots, n\}$, where $n = |\mathcal{V}|$ and tokens are ranked by descending probability.

3.2 Algorithm Overview

We formulate EES as a sequential stopping rule over the ranked token prefixes. Let

$$\kappa_\theta = \min \left(\{k \in \{2, \dots, n\} : \bar{H}_k < \theta P_k\} \cup \{n+1\} \right) \quad (5)$$

denote the first violation index, where the sentinel value $n+1$ represents the case in which no violation occurs. The retained candidate-set size is then

$$k_\theta^* = \min\{\kappa_\theta - 1, n\}. \quad (6)$$

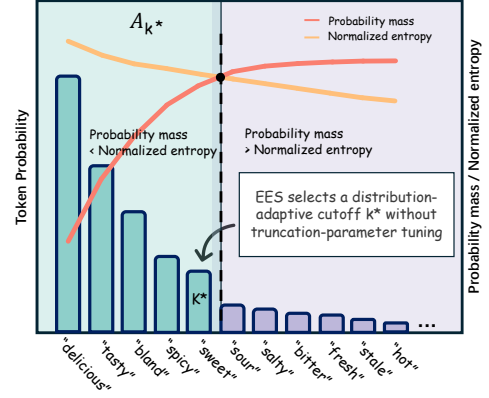


Figure 2: Mechanism of EES. The algorithm incrementally expands the candidate set and compares normalized entropy $\bar{H}_k$ with probability mass P_k at each step. If the first violation $\bar{H}_k < P_k$ occurs at index k , EES retains the preceding top- $(k-1)$ prefix; if no violation occurs, all n tokens are retained.

Why comparing $\bar{H}_k$ and P_k is meaningful. The two quantities characterize complementary properties of the same top- k candidate set:

- P_k measures **coverage**: the fraction of the full predictive probability captured by the top- k candidate set.
- $\bar{H}_k$ measures **relative uncertainty**: how dispersed the renormalized probabilities are within the top- k candidate set relative to a uniform distribution of the same size.

Both quantities are dimensionless, lie in $[0, 1]$, and are evaluated on the same ranked token prefix. We therefore use their relative ordering as a distribution-dependent signal for candidate-set expansion.

Setting θ . Equations 5 and 6 define a generalized stopping rule with a scaling factor $\theta > 0$. Throughout the main method and experiments, we fix $\theta = 1$ rather than tuning it for individual tasks or temperatures. The resulting EES rule is

$$\begin{aligned} \kappa &= \min \left(\{k \in \{2, \dots, n\} : \bar{H}_k < P_k\} \cup \{n+1\} \right), \\ k^* &= \min\{\kappa - 1, n\}. \end{aligned} \quad (7)$$

Appendix D reports the sensitivity analysis for θ , while all main results use the fixed value $\theta = 1$.

3.3 First-Crossing Criterion

Operationally, EES evaluates the condition $\bar{H}_k < P_k$ sequentially for $k = 2, \dots, n$ and stops at the first violation. It then retains the top- k^* prefix, where k^* is defined in Equation 7. Because the

stopping decision depends only on the earliest violation, subsequent values of $\bar{H}_k - P_k$ do not affect the output. This sequential procedure directly matches the implementation.

3.4 Algorithm Implementation

Figure 2 illustrates the sequential expansion and stopping mechanism of EES. The complete pseudocode is provided in Appendix A.

Complexity Analysis. The naive implementation has $\mathcal{O}(|\mathcal{V}| \log |\mathcal{V}|)$ complexity dominated by sorting, with $\mathcal{O}(k)$ per expansion for entropy computation ($\mathcal{O}(k^2)$ total for threshold selection). In Appendix B, we introduce an incremental entropy update that reduces per-expansion entropy computation from $\mathcal{O}(k)$ to $\mathcal{O}(1)$, yielding $\mathcal{O}(k^*)$ for threshold selection. Replacing sorting with a max-heap further reduces the overall complexity to $\mathcal{O}(|\mathcal{V}|)$ in theory; however, GPU-optimized sorting remains faster in practice, yielding a latency of $1.30\times$ relative to static methods.

4 Experiment

4.1 Setups

Models and Datasets We evaluated our method using Qwen2.5-7B, Qwen2.5-32B (Yang et al., 2024), and Llama3.1-8B (Grattafiori et al., 2024) on CommonsenseQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021). Detailed statistical information regarding all datasets is provided in Appendix E.

Baselines We compared against eight stochastic sampling methods: temperature sampling, top-p, top-k sampling, eta sampling, mirostat, typical sampling, adaptive sampling, and min-p (Nguyen et al., 2025); and two deterministic methods: Contrastive Search (CS) (Su et al., 2022) and Contrastive Decoding (CD) (Li et al., 2023).

Experimental Protocol We randomly sampled 20% of the training data as a validation set to identify optimal hyperparameters for each sampling method across different temperature settings. We configured the hyperparameter search space for each method based on recommendations from the original papers and subsequent literature (Shi et al., 2024):

- **Top-p:** $p \in \{0.75, 0.8, 0.85, 0.9, 0.95\}$
- **Top-k:** $k \in \{5, 10, 20, 50, 100\}$
- **Eta sampling:** $\eta \in \{3 \times 10^{-4}, \dots, 4 \times 10^{-3}\}$

- **Mirostat:** $\tau \in \{2.5, 3.0, 3.5, 4.0\}$
- **Typical sampling:** $p \in \{0.2, 0.9, 0.92, 0.95\}$
- **Adaptive sampling:** $\epsilon \in \{5 \times 10^{-4}, \dots, 1 \times 10^{-2}\}$
- **Min-p:** $p \in \{0.05, 0.1, 0.2\}$
- **CS:** $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$
- **CD:** $\alpha = 0.1, \beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

For CD, we use Qwen2.5-0.5B-Instruct and Qwen3-0.6B as student models for the Qwen2.5 and Qwen3 families, respectively. Since Llama3.1-8B lacks a smaller variant in the same family, CD is not evaluated on this model. EES requires no truncation-specific hyperparameter tuning or auxiliary models.

For each configuration, we conducted three independent runs and selected the hyperparameter setting that achieves the highest average accuracy on the validation set. On the test set, we evaluated each method using its optimal hyperparameter with five independent runs, reporting average performance as the final score. Following (Shi et al., 2024), we constructed model inputs by randomly sampling multiple examples from the training set to provide in-context learning examples.

4.2 QA Results

Table 1 shows that EES remains competitive across all configurations. While some baselines occasionally score slightly higher (e.g., Adaptive on Llama3.1-8B, Min-p on Qwen2.5-32B at $\tau=0.5$), these gains rely on temperature-specific hyperparameter tuning. In practical deployment where temperature may vary across tasks or users, EES offers a distinct advantage by maintaining competitive accuracy without retuning a truncation hyperparameter. In Appendix J, we show that fixing a baseline’s $\tau=0.8$ -optimal hyperparameter and applying it at other temperatures causes 4.6–7.2% accuracy drops, confirming the practical cost of temperature-specific tuning.

Notably, the deterministic baselines CS and CD consistently underperform EES on both QA datasets. Min-p exhibits model-dependent performance: it is competitive on Qwen2.5-32B but degrades significantly on Llama3.1-8B, where its optimal hyperparameter also varies across temperatures.

The tuned baselines require different optimal configurations across models and temperatures (e.g., top-k from 5 to 100, eta sampling from 3×10^{-4} to 4×10^{-3} , and Min-p from 0.05 to

LM	Sampling Method	CommonsenseQA				StrategyQA			
		$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$	Avg	$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$	Avg
Qwen2.5-7B	Temperature	83.70	82.39	82.01	82.70	75.75	75.90	74.67	75.44
	Top-p	83.77 _(0.75)	82.64 _(0.80)	81.93 _(0.80)	82.78	74.79 _(0.90)	75.87 _(0.85)	75.05 _(0.80)	75.24
	Top-k	82.75 ₍₁₀₎	82.42 ₍₁₀₀₎	82.65 ₍₂₀₎	82.61	75.87 ₍₅₎	75.11 ₍₅₎	74.44 ₍₁₀₀₎	75.14
	Eta	82.98 _(0.0009)	81.88 _(0.002)	82.10 _(0.0006)	82.32	75.43 _(0.0006)	74.99 _(0.0006)	75.11 _(0.002)	75.18
	Mirostat	83.88 _(2.5)	83.23 _(2.5)	83.55 _(2.5)	83.55	75.72 _(2.5)	76.33 _(3.5)	75.37 _(2.5)	75.81
	Typical	83.36 _(0.95)	84.11 _(0.2)	82.95 _(0.2)	83.47	75.34 _(0.2)	75.57 _(0.2)	75.23 _(0.2)	75.38
	Adaptive	83.62 _(0.0005)	83.46 _(0.005)	83.39 _(0.01)	83.49	75.63 _(0.005)	75.52 _(0.005)	75.46 _(0.005)	75.54
	Min-p	83.87 _(0.1)	83.16 _(0.1)	83.16 _(0.2)	83.40	75.02 _(0.1)	75.52 _(0.1)	75.02 _(0.1)	75.19
	CS	83.31	83.05	81.43	82.60	73.25	72.75	72.37	72.79
	CD	82.85	82.72	82.18	82.58	72.98	72.90	73.65	73.18
	Ours	84.42	<u>83.95</u>	83.64	84.00	76.33	76.56	75.95	76.28
Llama3.1-8B	Temperature	76.69	75.71	73.63	75.34	73.80	72.34	69.02	71.72
	Top-p	76.81 _(0.75)	76.25 _(0.75)	75.28 _(0.75)	76.11	73.54 _(0.95)	73.54 _(0.85)	72.66 _(0.75)	73.25
	Top-k	<u>77.33</u> ₍₂₀₎	76.23 ₍₅₎	74.76 ₍₅₎	76.11	75.28 ₍₅₎	72.20 ₍₂₀₎	72.52 ₍₅₎	73.33
	Eta	77.17 _(0.0006)	75.22 _(0.0006)	74.25 _(0.004)	75.55	74.06 _(0.002)	71.99 _(0.0003)	70.39 _(0.0009)	72.15
	Mirostat	77.20 _(2.5)	76.38 _(3.0)	76.54 _(2.5)	76.71	74.12 _(4.0)	75.46 _(3.0)	74.50 _(4.0)	74.69
	Typical	77.22 _(0.2)	75.81 _(0.2)	75.18 _(0.2)	76.07	75.05 _(0.2)	73.57 _(0.2)	72.02 _(0.90)	73.55
	Adaptive	77.17 _(0.01)	77.38 _(0.0005)	77.20 _(0.01)	77.25	75.11 _(0.0005)	<u>75.17</u> _(0.01)	75.14 _(0.0005)	75.14
	Min-p	71.76 _(0.2)	71.33 _(0.2)	70.73 _(0.05)	71.27	67.42 _(0.2)	67.71 _(0.2)	67.80 _(0.05)	67.64
	CS	71.70	70.38	67.96	70.01	67.57	67.13	63.23	65.98
	Ours	77.44	<u>76.72</u>	<u>76.90</u>	<u>77.02</u>	<u>75.14</u>	74.50	<u>74.61</u>	<u>74.75</u>
Qwen2.5-32B	Temperature	88.98	88.27	88.52	88.59	80.29	79.39	79.04	79.57
	Top-p	88.63 _(0.90)	88.21 _(0.95)	88.19 _(0.85)	88.34	80.44 _(0.80)	79.42 _(0.95)	79.24 _(0.80)	79.70
	Top-k	88.83 ₍₁₀₎	89.12 ₍₅₎	88.11 ₍₅₀₎	88.69	79.77 ₍₅₎	80.38 ₍₅₎	78.57 ₍₁₀₀₎	79.57
	Eta	88.99 _(0.002)	88.48 _(0.002)	88.47 _(0.002)	88.65	80.15 _(0.0009)	78.95 _(0.004)	79.39 _(0.004)	79.50
	Mirostat	89.24 _(3.0)	88.75 _(4.0)	89.09 _(2.5)	89.03	80.32 _(3.5)	80.12 _(2.5)	80.15 _(3.5)	80.20
	Typical	89.19 _(0.2)	<u>88.85</u> _(0.2)	88.70 _(0.92)	88.91	80.06 _(0.90)	80.32 _(0.92)	79.45 _(0.92)	79.94
	Adaptive	88.98 _(0.0005)	88.76 _(0.01)	88.83 _(0.001)	88.86	80.32 _(0.0005)	80.79 _(0.0005)	80.99 _(0.01)	80.70
	Min-p	89.39 _(0.05)	88.78 _(0.2)	88.40 _(0.1)	88.86	79.62 _(0.05)	<u>80.55</u> _(0.2)	79.91 _(0.1)	80.03
	CS	88.75	88.39	88.14	88.43	79.59	79.33	79.30	79.41
	CD	88.52	88.52	87.98	88.34	79.77	79.45	79.33	79.52
	Ours	<u>89.30</u>	<u>88.85</u>	<u>88.94</u>	89.03	<u>80.35</u>	80.52	<u>80.23</u>	<u>80.37</u>

Table 1: Cross-model performance comparison of sampling methods under different temperature settings on two QA datasets. Bold: best; underlined: second-best; parentheses: optimal hyperparameter. All values are averages over 5 independent runs; standard deviations range from 0.09 to 2.58 (mean 0.73); full per-cell standard deviations are in Appendix L.

0.2), whereas EES avoids temperature-specific re-tuning of a truncation hyperparameter.

4.3 Diversity analysis

We evaluated diversity using the repetition-based metric $Diversity = \prod_{n=2}^4 (1.0 - \frac{rep-n}{100})$ from (Meister et al., 2023) on correctly classified responses from Qwen2.5-7B. Figure 3 shows that EES achieves competitive diversity while maintaining strong accuracy. The accuracy-diversity trade-off is task-dependent: on CommonsenseQA, all methods reach high diversity ($>93\%$) at $\tau = 0.5$, with higher temperatures providing negligible diversity gains while degrading accuracy. On StrategyQA, EES maintains competitive accuracy but can yield lower diversity than some tuned baselines at comparable accuracy. This behavior is consistent with the smaller candidate sets selected by EES on this task, which favor exploitation when the predictive distribution is concentrated. Even at $\tau = 1.5$, EES maintains competitive accuracy without retuning a truncation hyperparameter.

4.4 Mathematical Reasoning

We evaluated mathematical reasoning performance on the GSM8K dataset (Cobbe et al., 2021) using Qwen3-8B (Yang et al., 2025). We examined two distinct generation modes: **w/o thinking** (standard direct prompting) and **w/ thinking** (reasoning mode), across temperature settings $\tau \in \{0.5, 0.8, 1.0\}$ to assess robustness. We adhered to the protocol in Section 4.1, using the same hyperparameter search pipeline. Final results represent the average of 5 independent runs. Each evaluation involves generating the full reasoning path and final answer, with correctness determined by exact match with the ground truth.

Results. Table 2 shows that EES ranks first or second in most configurations. In the **w/o thinking** mode, EES achieves the highest score at $\tau = 0.5$ and the second-best average accuracy (62.53). Min-p achieves the best single-temperature result at $\tau = 0.8$ (62.87) but is inconsistent across temperatures, with its optimal hyperparameter shifting from 0.2 at lower temperatures to 0.05 at $\tau = 1.0$. The deterministic baselines CS and CD lag behind,

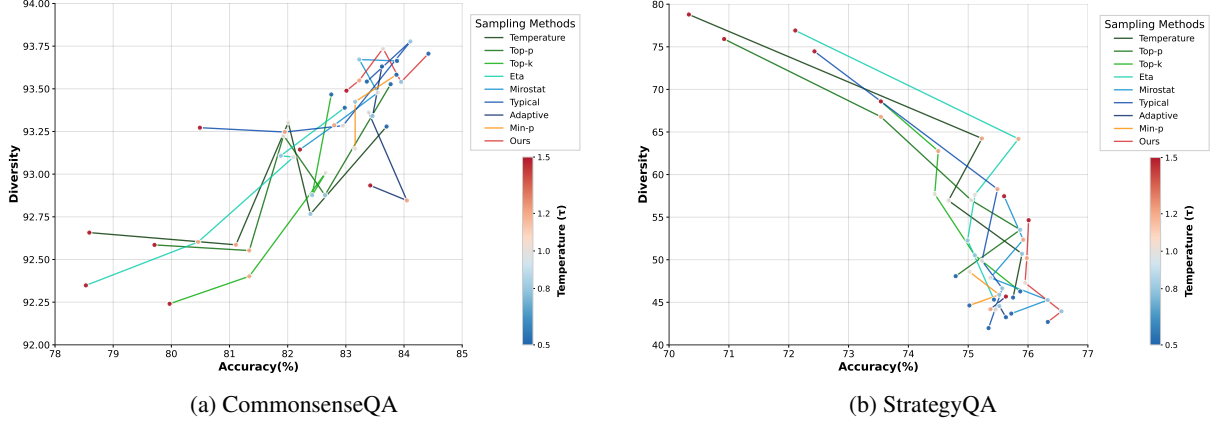


Figure 3: Accuracy-diversity performance across sampling methods and temperatures on two QA datasets using Qwen2.5-7B.

LM	Sampling Method	GSM8K (w/o thinking)				GSM8K (w/ thinking)			
		$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$	Avg	$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$	Avg
Qwen3-8B	Temperature	61.49	61.83	61.91	61.74	66.87	66.55	66.16	66.53
	Top-p	62.64 _(0.85)	62.38 _(0.95)	61.61 _(0.85)	62.21	66.94 _(0.90)	66.64 _(0.75)	66.31 _(0.80)	66.63
	Top-k	61.93 ₍₅₀₎	62.11 ₍₂₀₎	61.59 ₍₂₀₎	61.88	66.50 ₍₅₀₎	66.10 ₍₁₀₀₎	66.26 ₍₁₀₀₎	66.29
	Eta	62.58 _(0.004)	62.30 _(3e-4)	62.44 _(0.004)	62.44	66.72 _(0.002)	66.75 _(3e-4)	66.70 _(0.004)	66.72
	Mirostat	62.71 _(4.0)	62.52 _(4.0)	62.70 _(3.0)	62.64	66.61 _(3.0)	66.64 _(3.0)	66.55 _(4.0)	66.60
	Typical	62.56 _(0.2)	62.00 _(0.9)	62.09 _(0.9)	62.22	66.96 _(0.9)	66.35 _(0.2)	66.54 _(0.2)	66.62
	Adaptive	62.68 _(5e-4)	61.93 _(0.005)	62.08 _(0.001)	62.23	66.49 _(0.005)	66.73 _(5e-4)	66.22 _(0.01)	66.48
	Min-p	62.15 _(0.2)	62.87 _(0.2)	62.05 _(0.05)	62.36	66.84 _(0.1)	66.78 _(0.05)	66.75 _(0.2)	66.79
	CS	62.03	62.00	61.33	61.79	66.61	66.63	66.16	66.47
	CD	60.58	61.46	62.02	61.35	66.49	66.64	66.31	66.48
	Ours	62.85	62.55	62.18	62.53	66.64	66.76	66.75	66.72

Table 2: Comparison of sampling methods on GSM8K using Qwen3-8B, evaluating performance **w/o thinking** and **w/ thinking**. CS: Contrastive Search; CD: Contrastive Decoding. Values represent accuracy (%) averaged over 5 independent runs, with optimal hyperparameters in parentheses; standard deviations range from 0.14 to 0.85 (mean 0.47); full per-cell standard deviations are in Appendix L.

averaging 61.79 and 61.35 respectively compared to 62.53 for EES.

In the **w/ thinking** mode, Min-p achieves the highest average (66.79), followed by EES and Eta (both 66.72). However, Min-p requires temperature-specific tuning (hyperparameters vary across τ), while EES maintains consistently high performance (66.64–66.76%) without adjustment. CS and CD perform comparably to temperature sampling (~ 66.48 avg), confirming that deterministic scoring provides limited benefit for extended reasoning.

4.5 Creative Writing Evaluation

To assess the generation quality of our proposed method in open-ended text generation, we conducted experiments on the WikiText-103 dataset (Merity et al., 2016). We utilized the Qwen2.5-7B and Llama3.1-8B models, with generation performed at an elevated temperature of $\tau = 1.0$ to promote diversity. Each generational

task utilized 32-word prefixes as prompts, with model outputs limited to 256 tokens. All hyperparameters of baseline methods were optimized on a 200-sample validation set by maximizing the MAUVE score (Pillutla et al., 2021) against human references.

Instead of relying on automatic metrics, which often fail to align with human preferences (Nimah et al., 2023), we adopted a pairwise comparison framework (Zhu et al., 2024) involving both human annotators and an LLM ensemble as judges. Following (Nguyen et al., 2025), we conducted evaluations across two key dimensions: (1) **Quality**, covering coherence, relevance to the prompt, and overall fluency; and (2) **Diversity**, assessing the creativity and distinctiveness of the content.

Evaluation Protocol. For human evaluation, we used generations from Qwen2.5-7B and recruited 10 diverse, fluent English speakers. Each participant evaluated 500 text pairs (50 per baseline) on a 3-point Likert scale: +1 indicates our method

Comparison	Human Evaluators (%)			LLM Ensemble (%)		
	Win	Tie	Loss	Win	Tie	Loss
<i>Quality Evaluation</i>						
vs Human Ref.	9.20	68.40	22.40	2.00	2.00	96.00
vs Top-p	24.60	54.20	21.20	44.50	7.50	48.00
vs Top-k	23.80	57.60	18.60	46.00	9.50	44.50
vs Eta	44.20	38.40	17.40	48.00	7.00	45.00
vs Mirostat	25.40	58.20	16.40	53.50	17.50	29.00
vs Typical	34.20	40.60	25.20	61.00	5.50	33.50
vs Adaptive	38.40	52.80	8.80	56.50	18.00	25.50
vs Min-p	38.00	46.00	16.00	54.00	16.00	30.00
vs CS	38.40	48.20	13.40	52.00	8.00	40.00
vs CD	35.20	50.80	14.00	54.50	12.50	33.00
<i>Diversity Evaluation</i>						
vs Human Ref.	14.80	30.20	55.00	3.00	9.00	88.00
vs Top-p	21.60	31.40	47.00	20.00	22.00	58.00
vs Top-k	28.40	27.60	44.00	31.00	20.50	48.50
vs Eta	42.80	33.20	24.00	25.00	24.50	50.50
vs Mirostat	38.60	26.40	35.00	40.50	29.50	30.00
vs Typical	32.40	18.60	49.00	32.50	25.00	42.50
vs Adaptive	46.20	35.40	18.40	51.00	30.00	19.00
vs Min-p	52.00	30.00	18.00	58.00	24.00	18.00
vs CS	32.40	28.60	39.00	35.00	26.00	39.00
vs CD	28.60	31.40	40.00	36.50	27.50	36.00

Table 3: Pairwise comparison results from Human evaluators and LLM judges.

is clearly superior, -1 indicates the baseline is superior, and 0 indicates comparable quality. We excluded temperature-only sampling because pilot evaluations showed it produced noticeably incoherent outputs at $\tau = 1.0$ (all 5 pilot annotators rated it lowest), making pairwise comparisons uninformative; we instead included CS, CD, and human-written reference texts from WikiText-103 as benchmarks. Additional details on the evaluation interface, compensation, and inter-annotator agreement are provided in Appendix H.

For automated evaluation, we employed a suite of state-of-the-art LLMs as judges, including Claude Sonnet 4 (Anthropic, 2025a), Claude Sonnet 4.5 (Anthropic, 2025b), GPT-4.1 (OpenAI, 2025), and Gemini 2.5 Pro (Google DeepMind, 2025). We utilized carefully designed prompts following (Nguyen et al., 2025) to leverage their demonstrated capability in aligning with human judgment.

Results. Table 3 summarizes the findings. In **Quality**, EES secures positive win-loss ratios against most baselines, with clear advantages over deterministic methods (38.4% vs. 13.4% against CS; 35.2% vs. 14.0% against CD). The LLM ensemble corroborates this trend. In **Diversity**, EES is preferred over Adaptive and Min-p, whereas CS and CD receive slightly higher diversity preferences but lower quality preferences. Against human references, a quality gap remains (96% LLM

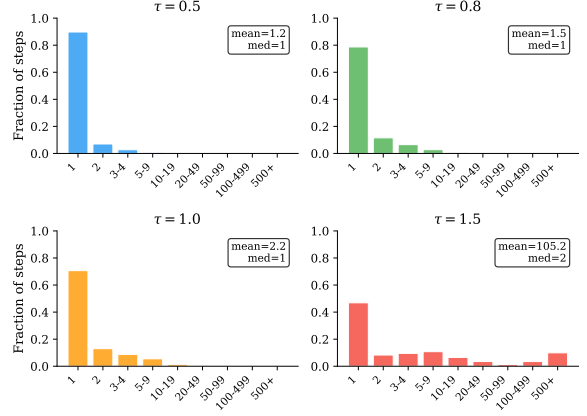


Figure 4: Distribution of k^* values selected by EES at different temperatures on CommonsenseQA (Qwen2.5-7B-Instruct). At $\tau = 0.5$, over 89% of steps select $k^* = 1$; at $\tau = 1.5$, the distribution becomes heavy-tailed with a substantial fraction of steps retaining >100 candidates.

	$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$	$\tau = 1.5$
Mean k^*	1.2	1.5	2.2	105.2
Median k^*	1	1	1	2
Mean P_κ	0.919	0.880	0.848	0.817
Mean $\bar{H}_\kappa$	0.287	0.387	0.442	0.543

Table 4: Statistics of the retained candidate-set size k^* and the quantities evaluated at the first-violation index κ at different temperatures on CommonsenseQA (Qwen2.5-7B-Instruct, 200 samples).

loss rate), though human evaluators often found EES comparable (68.4% tie).

5 Analysis

The preceding experiments demonstrate *that* EES achieves robust performance across temperatures without retuning a truncation hyperparameter. In this section, we investigate *why*.

5.1 How EES Adapts to Temperature

A key claim of our method is that the retained candidate-set size k^* automatically adapts to the shape of the probability distribution, which changes with temperature. To verify this, we record the value of k^* at every generation step during inference on CommonsenseQA using Qwen2.5-7B across temperatures $\tau \in \{0.5, 0.8, 1.0, 1.5\}$.

Figure 4 and Table 4 reveal the adaptive mechanism. At low temperature ($\tau = 0.5$), the mean k^* is just 1.2 with over 89% of steps selecting $k^* = 1$ (near-greedy), and P_κ averages 0.919 at the corresponding first-violation indices. As temperature

increases, k^* grows: at $\tau = 1.5$, the mean reaches 105.2 with a heavy-tailed distribution, while $\bar{H}_\kappa$ increases from 0.287 to 0.543, indicating that the first violation occurs at prefixes with higher normalized entropy. Conversely, P_κ decreases from 0.919 to 0.817, reflecting that more probability mass is spread across the tail. This is exactly the behavior practitioners achieve manually by increasing top- p at higher temperatures—but EES does it automatically.

This reveals the core mechanism: k^* **reflects the temperature’s effect on the token-level probability distribution**. Rather than retuning a truncation hyperparameter for each temperature, EES applies the same first-crossing rule at every generation step, allowing the retained candidate-set size to vary with the distribution. We provide additional analysis of step-level truncation behavior and model scale effects in Appendix F.

6 Related Work

Deterministic Methods. Greedy sampling and beam search (Freitag and Al-Onaizan, 2017) provide coherent but repetitive outputs (Vijayakumar et al., 2016). Contrastive Search (Su et al., 2022) balances model confidence with lexical diversity penalties, and Contrastive Decoding (Li et al., 2023) uses the difference between a large and small model’s logits. DoLa (Chuang et al., 2024) leverages layer-wise representational differences. These methods modify the scoring function to improve generation quality, whereas EES addresses the truncation problem in stochastic sampling—determining which tokens to *include* in the candidate set before probabilistic selection. Our experiments show that EES achieves higher task accuracy than CS and CD where these methods are evaluated.

Stochastic Methods. Top- k (Fan et al., 2018) and top- p (Holtzman et al., 2019) employ fixed truncation criteria based on candidate set size and probability mass, respectively. Recent adaptive methods dynamically adjust candidate sets: η -sampling (Hewitt et al., 2022) uses entropy thresholds, Mirostat (Basu et al., 2020) maintains perplexity within a target range, typical sampling (Meister et al., 2023) ranks tokens by divergence from the distribution’s entropy, adaptive sampling (Zhu et al., 2024) uses entropy-based confidence metrics, min- p (Nguyen et al., 2025) scales thresholds by the top token’s probability. Arias et al. (2025) pro-

vide a systematic study of how decoding hyperparameters interact; Renze (2024) show that temperature changes within $[0, 1]$ have limited impact on problem-solving tasks. These methods retain method-specific control parameters, whereas EES uses a fixed first-crossing rule to determine the retained candidate-set size without tuning a truncation-specific parameter.

Relation to Entropy-Aware Methods. Mirostat (Basu et al., 2020) controls generation toward a user-specified target perplexity, typical sampling (Meister et al., 2023) ranks tokens by the deviation of their information content from the distribution entropy, and η -sampling (Hewitt et al., 2022) applies an entropy-dependent probability threshold. Entropy-aware Dynamic Temperature (EDT) (Zhang et al., 2024) instead uses entropy to adapt the temperature rather than truncating the candidate set. EES differs from these methods by comparing normalized entropy with cumulative probability mass over ranked token prefixes and stopping expansion at the first violation, retaining the preceding prefix. It therefore determines the candidate-set size without a user-specified truncation target.

7 Conclusion

We presented Entropy Equilibrium Sampling (EES), a sampling strategy that determines the retained candidate-set size using a sequential first-crossing rule that compares normalized entropy with cumulative probability mass over ranked token prefixes, without requiring a user-tuned truncation threshold. EES automatically adapts its truncation behavior—near-greedy under peaked distributions and broader under flat ones—without retuning a truncation hyperparameter across temperature settings. Evaluations across commonsense, strategic, and mathematical reasoning as well as open-ended generation show that EES remains competitive with tuned baselines while maintaining diversity. The retained candidate-set size k^* reflects temperature-dependent changes in the probability distribution, increasing from a mean of 1.2 at $\tau = 0.5$ to 105.2 at $\tau = 1.5$. Although tuned baselines occasionally perform better in individual configurations, EES remains competitive without temperature-specific retuning of a truncation hyperparameter.

Future work includes optimizing inference latency via custom GPU kernels and extending EES to multimodal generation.

Limitations

We identify the following limitations of our method:

Inference Latency. Although the theoretical time complexity of EES can be reduced to $\mathcal{O}(|\mathcal{V}|)$ (as detailed in Appendix B), the practical runtime overhead remains a challenge. Current GPU implementations for dynamic thresholding operations (such as iterative summation or heap-based selection) are not as highly optimized as the mature kernels used for static Top- k or Top- p sampling. Consequently, despite the theoretical efficiency, the actual inference latency is higher than that of static methods, which may impact throughput in latency-sensitive applications.

Safety Considerations. Because EES can retain larger candidate sets under flatter distributions, it may increase the probability of sampling harmful or toxic tokens when the model assigns non-negligible probability to such outputs. Unlike methods controlled by a fixed top- k size or top- p mass threshold, EES determines the retained candidate set through the first-crossing rule applied to the current probability distribution. In safety-critical applications, combining EES with output filtering or safety-aligned decoding constraints is recommended.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No.62276154); the Natural Science Foundation of Guangdong Province (Grant No.2024TQ08X729); Basic Research Fund of Shenzhen City (Grant No.JCYJ20240813112009013 and GJHZ20240218113603006); The Major Key Project of PCL for Experiments and Applications (Grant No.PCL2024A08).

References

- Anthropic. 2025a. Claude 4 technical report. <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>.
- Anthropic. 2025b. Claude Sonnet 4.5 system card. <https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf>.
- Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020.
- Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R Varshney. 2020. Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint arXiv:2007.14966*.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

- Google DeepMind. 2025. Gemini 2.5 Pro model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- John Hewitt, Christopher D Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Lizhi Liao, Heng Li, Weiyi Shang, and Lei Ma. 2022. An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(3):1–40.
- Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5328–5339.
- 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. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *International Conference on Learning Representations*, volume 2025, pages 70333–70366.
- Ifritahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266.
- OpenAI. 2025. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>.
- 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. *Advances in Neural Information Processing Systems*, 34:4816–4828.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7346–7356.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629.
- Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.
- Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models. *arXiv preprint arXiv:2406.16838*.
- Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.
- Peng Xing, Yinghui Li, Shirong Ma, Xinnian Liang, Haojing Huang, Yangning Li, Hai-Tao Zheng, Wenhao Jiang, and Ying Shen. 2024. Mitigating catastrophic forgetting in multi-domain chinese spelling correction by multi-stage knowledge transfer framework. *arXiv preprint arXiv:2402.11422*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. [Qwen2.5 technical report](#). *ArXiv*, abs/2412.15115.
- Shaoxiong Zhan, Yanlin Lai, Ziyu Lu, Dahua Lin, Ziqing Yang, and Fei Tan. 2026. Mathsmith: Towards extremely hard mathematical reasoning by forging synthetic problems with a reinforced policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34602–34610.
- Shaoxiong Zhan, Hai Lin, Hongming Tan, Xiaodong Cai, Hai-Tao Zheng, Xin Su, Zifei Shan, Ruitong Liu, and Hong-Gee Kim. 2025. Lexsembridge: Fine-grained dense representation enhancement through token-aware embedding augmentation. *arXiv preprint arXiv:2508.17858*.
- Shimao Zhang, Yu Bao, and Shujian Huang. 2024. Edt: Improving large language models’ generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*.
- Wenhong Zhu, Hongkun Hao, Zhiwei He, Yiming Ai, and Rui Wang. 2024. Improving open-ended text generation via adaptive decoding. *arXiv preprint arXiv:2402.18223*.

A Algorithm Detail

Algorithm 1 presents the complete pseudocode for EES.

Algorithm 1 Entropy Equilibrium Sampling (EES)

Require: Context $\mathcal{C}$; Language Model P_{LM}

Ensure: Generated Sequence $\mathcal{G}$

```

1:  $\mathcal{G} \leftarrow \emptyset$ 
2: for each generation step do
3:    $\mathbf{p} \leftarrow \text{Sort}(P_{\text{LM}}(\cdot|\mathcal{C}), \text{descending})$ 
4:    $P_0 \leftarrow 0, \kappa \leftarrow |\mathcal{V}| + 1$ 
5:   for  $k = 1$  to  $|\mathcal{V}|$  do
6:      $P_k \leftarrow P_{k-1} + p_k$ 
7:     if  $k \geq 2$  then
8:        $\bar{H}_k \leftarrow \frac{-\sum_{i=1}^k (p_i/P_k) \log(p_i/P_k)}{\log k}$ 
9:       if  $\bar{H}_k < P_k$  then
10:         $\kappa \leftarrow k$ ; break
11:      end if
12:    end if
13:  end for
14:   $k^* \leftarrow \min\{\kappa - 1, |\mathcal{V}|\}$ 
15:   $x_t \leftarrow \text{sample from top-}k^* \text{ distribution}$ 
16:  Update  $\mathcal{C}$  and  $\mathcal{G}$  with  $x_t$ 
17: end for
18: return  $\mathcal{G}$ 

```

B Incremental Entropy Computation

While Section 3.4 presents the conceptual framework, practical implementation can avoid redundant entropy computations through incremental updates.

Incremental Entropy Update: The entropy H_{k+1} for the expanded candidate set can be computed as:

$$H_{k+1} = \frac{P_k}{P_{k+1}} H_k + \frac{P_k}{P_{k+1}} \log \frac{P_{k+1}}{P_k} - \frac{p_{k+1}}{P_{k+1}} \log \frac{p_{k+1}}{P_{k+1}} \quad (8)$$

where $P_{k+1} = P_k + p_{k+1}$. This reduces the per-expansion entropy computation from $\mathcal{O}(k)$ to $\mathcal{O}(1)$, yielding $\mathcal{O}(k^*)$ for threshold selection. Combined with a max-heap (which avoids the $\mathcal{O}(|\mathcal{V}| \log |\mathcal{V}|)$ sorting step), the overall complexity becomes $\mathcal{O}(|\mathcal{V}|)$. However, in practice GPU-optimized sorting outperforms heap-based extraction, so our implementation retains sorting.

Proof sketch: By writing $p_i^{(k+1)} = p_i^{(k)} \cdot \frac{P_k}{P_{k+1}}$ for the first k tokens and expanding the entropy

sum, the cross-terms simplify to the stated formula. The full derivation follows by expanding the renormalized entropy after adding the $(k+1)$ -st token.

Computational Efficiency. Table 5 compares inference latency across methods on Qwen3-8B (batch=128, max 128 tokens).

Method	Avg. Latency (ms/step)	Sampling (ms/step)	Relative Slowdown
Temperature	69.65	0.02	1.00×
Top- p	71.54	0.01	1.03×
Top- k	69.38	0.78	1.00×
Typical	72.25	1.26	1.04×
Adaptive	118.11	75.25	1.70×
EES (ours)	90.59	47.42	1.30×

Table 5: Inference latency comparison. EES introduces moderate overhead (1.30×) while being 23% faster than Adaptive Sampling.

C Justification for $\theta = 1$

We provide the detailed argument for why $\theta = 1$ is the natural principled default.

Symmetry. Both $\bar{H}_k$ and P_k are normalized to $[0, 1]$ by construction: $\bar{H}_k$ normalizes Shannon entropy by $\log k$, and P_k is a cumulative probability. Because they already share the same scale, comparing them directly ($\theta = 1$) introduces no prior bias toward either measure. Setting $\theta > 1$ implicitly prioritizes probability mass (exploitation); setting $\theta < 1$ implicitly prioritizes entropy (exploration). $\theta = 1$ lets the truncation decision be determined entirely by the distribution itself. While nearby values (e.g., $\theta \in [0.9, 1.2]$) yield similar empirical performance (Table 6), $\theta = 1$ is the natural bias-free operating point.

Sequential interpretation. For $\theta = 1$, EES scans the ranked token prefixes in increasing order of k . It stops at the first index κ satisfying $\bar{H}_\kappa < P_\kappa$ and retains $k^* = \min\{\kappa - 1, n\}$. Only this first violation determines the selected candidate set.

D Ablation Study on θ

We conducted comprehensive ablation experiments across $\theta \in \{0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4\}$ on multiple models and tasks, evaluating at three temperature settings ($\tau \in \{0.5, 0.8, 1.0\}$) and reporting averaged results.

LM	Task	$\theta=0.6$	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
Qwen2.5-7B	CSQA	82.50	82.95	83.33	83.72	84.00	83.92	83.96	83.99	84.01
	SQA	75.53	75.62	75.67	75.94	76.28	76.15	76.23	76.24	76.22
Qwen2.5-32B	CSQA	88.82	88.70	88.69	88.69	89.03	88.85	88.93	88.90	88.91
	SQA	79.46	79.66	79.72	79.90	80.37	80.35	80.39	80.31	80.31
Llama3.1-8B	CSQA	76.81	76.57	76.78	76.84	77.02	77.10	76.91	77.03	76.99
	SQA	75.30	75.12	74.95	74.84	74.75	75.03	74.49	74.34	74.39
Qwen3-8B	GSM8K (w/o)	61.76	62.03	62.03	62.31	62.53	62.37	62.55	62.48	62.15
	GSM8K (w/)	66.15	66.60	66.40	66.56	66.72	66.86	66.71	66.81	66.66

Table 6: Accuracy (%) vs. θ , averaged across $\tau \in \{0.5, 0.8, 1.0\}$. CSQA: CommonsenseQA; SQA: StrategyQA. Bold: best per row.

Analysis. Table 6 shows that performance remains broadly stable around $\theta = 1$, although the best value varies across models and tasks. Decreasing θ tends to expand the candidate set, whereas increasing θ tends to reduce it. The fixed choice $\theta = 1$ lies within a competitive neighborhood and is used across all experiments without per-task or per-temperature retuning.

E Dataset Statistics

Dataset	Task Type	Train	Test
CommonsenseQA	Multiple Choice QA	9,741	1,221
StrategyQA	Boolean QA	2,290	490
WikiText-103	Open-Ended Generation	28,475	60
GSM8K	Math Reasoning	7,473	1,319

Table 7: Dataset statistics.

We randomly sampled 20% of training data as a validation set for hyperparameter tuning. For WikiText-103, we filtered articles >256 tokens and used 200 samples for validation, with MAUVE score as the optimization target.

F Additional Analysis

F.1 Truncation Behavior on Individual Steps

To illustrate step-level behavior, we examine representative generation steps from three qualitatively different distribution shapes.

Method	High-conf. ($H=0.02$)	Moderate ($H=0.68$)	High-unc. ($H=3.16$)
Top- p ($p=0.9$)	1	2	76
Top- k ($k=20$)	20	20	20
Eta ($\eta=0.001$)	2	6	46
EES (ours)	1	1	5

Table 8: Number of tokens retained by each method on three representative generation steps (Qwen2.5-7B-Instruct, $\tau=1.0$). EES provides conservative, entropy-aware truncation.

As shown in Table 8, top- k is entirely static. Top- p and eta adapt but can over-retain on uncertain steps (76 and 46 tokens). EES applies conservative truncation: $k^* = 1$ on confident and moderate steps (where the top token holds $>70\%$ of mass), expanding to only 5 on the high-uncertainty step. The critical distinction from greedy decoding is that EES *does* expand when genuinely needed.

F.2 Effect of Model Scale

Model	Mean H	Std H	Median H
Qwen2.5-7B	1.42	1.60	0.96
Llama3.1-8B	1.58	1.86	0.84
Qwen2.5-32B	1.60	1.58	1.25

Table 9: Next-token entropy statistics ($\tau = 1.0$) on CommonsenseQA.

While mean entropy is similar across scales, Llama3.1-8B has the highest entropy standard deviation (1.86), indicating greater variation in entropy across generation steps. Qwen2.5-32B has the lowest standard deviation (1.58) and a higher median entropy (1.25), indicating slightly more stable entropy values across generation steps. **Across the evaluated models, EES shows a larger relative performance advantage on the 7B/8B models, while the gap narrows on Qwen2.5-32B.**

G Hyperparameter Sensitivity

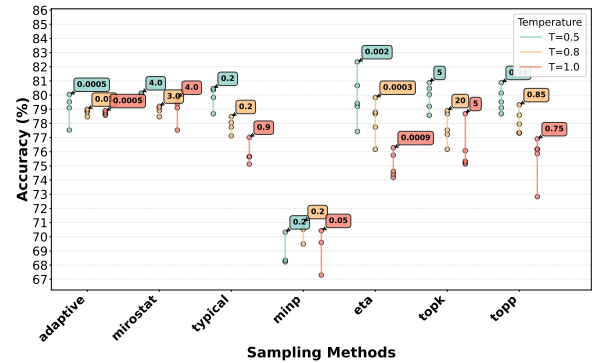


Figure 5: Performance variance of baseline methods under different hyperparameter settings at fixed temperatures. Methods with wider spread are more sensitive to hyperparameter choice.

Figure 5 demonstrates that existing methods exhibit substantial performance variance across hyperparameter configurations, with some configurations degrading accuracy by over 5%. This motivates methods that avoid temperature-specific retuning of truncation hyperparameters.

H Human Evaluation Protocol

Evaluators were presented with pairwise comparisons via a web interface displaying the prompt alongside two anonymized generations, as shown in Figure 6. They assessed outputs on Quality (coherence, relevance, fluency) and Diversity (creativity, distinctiveness) using a 3-point scale (+1/0/−1). We recruited 10 fluent English speakers; each evaluated 500 pairs. Evaluators were compensated at \$15/hour, above local minimum wage. Inter-annotator agreement (Fleiss’ κ) was 0.42 (moderate).

Figure 6: Screenshot of the human evaluation interface used for pairwise comparison.

I Statistical Significance

To rigorously assess performance differences, we conducted Welch’s t-tests comparing EES against each baseline method. Per-temperature p-values were combined using Fisher’s method and corrected with Bonferroni adjustment for multiple comparisons.

Key findings. EES shows significant improvements ($p < 0.05$) over static methods (Temperature, Top- p , Top- k) and deterministic baselines (CS, CD) on 7B/8B models. Against adaptive methods (Mirostat, Adaptive), differences are generally not significant, indicating that EES achieves comparable performance without retuning a truncation

Method	Setting	$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$
Top- k	Per- τ optimal	77.33 _(k=20)	76.23 _(k=5)	74.76 _(k=5)
	Fixed@ $\tau=0.8$	71.33 _(k=5)	76.23	—
	Δ	-6.00	0.00	—
Eta	Per- τ optimal	77.17 _($\eta=6e-4$)	75.22 _($\eta=6e-4$)	74.25 _($\eta=4e-3$)
	Fixed@ $\tau=0.8$	71.84 _($\eta=6e-4$)	75.22	67.01 _($\eta=6e-4$)
	Δ	-5.32	0.00	-7.24
Adaptive	Per- τ optimal	77.17 _($\epsilon=0.01$)	77.38 _($\epsilon=5e-4$)	77.20 _($\epsilon=0.01$)
	Fixed@ $\tau=0.8$	72.32 _($\epsilon=5e-4$)	77.38	72.61 _($\epsilon=5e-4$)
	Δ	-4.85	0.00	-4.59
EES (ours)	(fixed rule)	77.44	76.72	76.90

Table 10: Hyperparameter transfer across temperatures on CommonsenseQA (Llama3.1-8B). “Per- τ optimal” uses the best hyperparameter tuned at each temperature; “Fixed@ $\tau=0.8$ ” fixes the $\tau=0.8$ -optimal value. Δ is the accuracy drop. “—” indicates that the fixed and per- τ optimal hyperparameters coincide, so no transfer degradation occurs. Baseline methods degrade by 4.6–7.2% when transferred, while EES applies the same fixed rule across temperatures.

hyperparameter. On Qwen2.5-32B, most differences are non-significant. As shown in Table 9, this model also exhibits the lowest entropy variability across generation steps, which may help explain the narrower performance gap between EES and the baselines.

J Hyperparameter Transfer Across Temperatures

A key practical concern is whether hyperparameters tuned at one temperature transfer to other settings. We fix each method’s optimal hyperparameter from $\tau = 0.8$ and evaluate at $\tau = 0.5$ and $\tau = 1.0$ on CommonsenseQA using Llama3.1-8B (5 runs each).

Table 10 shows that all baseline methods suffer substantial accuracy degradation (4.6–7.2%) when their hyperparameters are transferred across temperatures. Eta sampling is especially sensitive, dropping 7.24% at $\tau=1.0$ because the entropy threshold tuned for lower temperatures becomes overly restrictive at higher temperatures. EES avoids this failure mode entirely by deriving the truncation threshold directly from the distribution at each step.

K Licensing Information

All datasets and models are publicly available: GSM8K (MIT), WikiText-103 (CC BY-SA 3.0), CommonsenseQA (MIT), StrategyQA (Apache 2.0), Qwen2.5/Qwen3 (Apache 2.0), Llama3.1 (Llama 3 Community License).

Baseline	CommonsenseQA			StrategyQA			GSM8K
	Qwen2.5-7B	Llama3.1-8B	Qwen2.5-32B	Qwen2.5-7B	Llama3.1-8B	Qwen2.5-32B	Qwen3-8B
Temperature	<.001**	<.001**	.022*	.139	.003**	.200	.030*
Top- <i>p</i>	<.001**	.005**	.002**	.044*	.008**	.652	.958
Top- <i>k</i>	<.001**	.232	.266	.040*	.161	.266	.042*
Eta	<.001**	<.001**	.067	.042*	<.001**	.042*	1.00
Mirostat	.302	1.00	1.00	1.00	1.00	1.00	1.00
Typical	.026*	.007**	1.00	.040*	.073	1.00	1.00
Adaptive	.006**	1.00	1.00	.403	1.00	1.00	1.00
Min- <i>p</i>	.030*	<.001**	1.00	.031*	<.001**	1.00	1.00
CS	<.001**	<.001**	.013*	<.001**	<.001**	.036*	.016*
CD	.001**	—	.005**	<.001**	—	.168	.001**

* $p < 0.05$, ** $p < 0.01$. “—”: CD not evaluated on Llama3.1-8B (no same-family small model).

Table 11: Statistical significance (Bonferroni-corrected p -values) of EES vs. baseline methods. Bold indicates $p < 0.05$. P -values are computed per temperature via Welch’s t -test, combined with Fisher’s method, and Bonferroni-corrected. For every statistically significant comparison, EES achieves higher mean accuracy than the corresponding baseline.

L Full Standard Deviation Tables

LM	Method	CommonsenseQA			StrategyQA		
		$\tau=0.5$	$\tau=0.8$	$\tau=1.0$	$\tau=0.5$	$\tau=0.8$	$\tau=1.0$
Qwen2.5-7B	Temperature	83.70 $\pm$ 1.05	82.39 $\pm$ 0.74	82.01 $\pm$ 0.50	75.75 $\pm$ 0.62	75.90 $\pm$ 1.21	74.67 $\pm$ 0.74
	Top- p	83.77 $\pm$ 0.56	82.64 $\pm$ 0.26	81.93 $\pm$ 0.41	74.79 $\pm$ 0.83	75.87 $\pm$ 0.72	75.05 $\pm$ 1.56
	Top- k	82.75 $\pm$ 0.48	82.42 $\pm$ 0.44	82.65 $\pm$ 0.58	75.87 $\pm$ 0.72	75.11 $\pm$ 1.47	74.44 $\pm$ 0.61
	Eta	82.98 $\pm$ 0.56	81.88 $\pm$ 0.80	82.10 $\pm$ 0.55	75.43 $\pm$ 1.10	74.99 $\pm$ 0.95	75.11 $\pm$ 0.91
	Mirostat	83.88 $\pm$ 0.49	83.23 $\pm$ 0.82	83.55 $\pm$ 0.63	75.72 $\pm$ 0.98	76.33 $\pm$ 0.75	75.37 $\pm$ 0.89
	Typical	83.36 $\pm$ 0.50	84.11 $\pm$ 0.25	82.95 $\pm$ 0.52	75.34 $\pm$ 0.77	75.57 $\pm$ 0.64	75.23 $\pm$ 0.56
	Adaptive	83.62 $\pm$ 0.15	83.46 $\pm$ 0.61	83.39 $\pm$ 0.33	75.63 $\pm$ 0.92	75.52 $\pm$ 1.14	75.46 $\pm$ 0.85
	Min- p	83.87 $\pm$ 0.51	83.16 $\pm$ 1.02	83.16 $\pm$ 0.31	75.02 $\pm$ 0.89	75.52 $\pm$ 0.80	75.02 $\pm$ 1.21
	Ours	84.42 $\pm$ 0.29	83.95 $\pm$ 0.38	83.64 $\pm$ 0.27	76.33 $\pm$ 0.66	76.56 $\pm$ 0.72	75.95 $\pm$ 0.92
	CS	83.31 $\pm$ 0.39	83.05 $\pm$ 0.38	81.43 $\pm$ 0.50	73.25 $\pm$ 0.49	72.75 $\pm$ 0.56	72.37 $\pm$ 1.32
	CD	82.85 $\pm$ 0.91	82.72 $\pm$ 0.91	82.18 $\pm$ 0.69	72.98 $\pm$ 0.73	72.90 $\pm$ 0.96	73.65 $\pm$ 0.44
Llama3.1-8B	Temperature	76.69 $\pm$ 0.60	75.71 $\pm$ 0.74	73.63 $\pm$ 0.82	73.80 $\pm$ 1.02	72.34 $\pm$ 1.24	69.02 $\pm$ 2.58
	Top- p	76.81 $\pm$ 0.09	76.25 $\pm$ 0.46	75.28 $\pm$ 0.88	73.54 $\pm$ 1.17	73.54 $\pm$ 0.81	72.66 $\pm$ 0.83
	Top- k	77.33 $\pm$ 0.36	76.23 $\pm$ 0.94	74.76 $\pm$ 1.43	75.28 $\pm$ 1.05	72.20 $\pm$ 1.16	72.52 $\pm$ 1.57
	Eta	77.17 $\pm$ 0.67	75.22 $\pm$ 0.67	74.25 $\pm$ 0.72	74.06 $\pm$ 1.48	71.99 $\pm$ 0.94	70.39 $\pm$ 0.91
	Mirostat	77.20 $\pm$ 0.44	76.38 $\pm$ 0.51	76.54 $\pm$ 0.59	74.12 $\pm$ 1.21	75.46 $\pm$ 1.18	74.50 $\pm$ 1.70
	Typical	77.22 $\pm$ 0.27	75.81 $\pm$ 0.33	75.18 $\pm$ 0.38	75.05 $\pm$ 0.86	73.57 $\pm$ 0.75	72.02 $\pm$ 1.11
	Adaptive	77.17 $\pm$ 0.21	77.38 $\pm$ 0.30	77.20 $\pm$ 0.09	75.11 $\pm$ 0.81	75.17 $\pm$ 0.56	75.14 $\pm$ 0.98
	Min- p	71.76 $\pm$ 0.48	71.33 $\pm$ 0.89	70.73 $\pm$ 1.11	67.42 $\pm$ 1.17	67.71 $\pm$ 1.19	67.80 $\pm$ 2.10
	Ours	77.44 $\pm$ 0.35	76.72 $\pm$ 0.63	76.90 $\pm$ 0.88	75.14 $\pm$ 0.58	74.50 $\pm$ 2.16	74.61 $\pm$ 0.89
	CS	71.70 $\pm$ 0.63	70.38 $\pm$ 1.47	67.96 $\pm$ 1.37	67.57 $\pm$ 1.09	67.13 $\pm$ 0.86	63.23 $\pm$ 1.64
	CD	71.70 $\pm$ 0.63	70.38 $\pm$ 1.47	67.96 $\pm$ 1.37	67.57 $\pm$ 1.09	67.13 $\pm$ 0.86	63.23 $\pm$ 1.64
Qwen2.5-32B	Temperature	88.98 $\pm$ 0.23	88.27 $\pm$ 0.16	88.52 $\pm$ 0.29	80.29 $\pm$ 0.85	79.39 $\pm$ 0.77	79.04 $\pm$ 1.12
	Top- p	88.63 $\pm$ 0.29	88.21 $\pm$ 0.46	88.19 $\pm$ 0.38	80.44 $\pm$ 0.68	79.42 $\pm$ 1.01	79.24 $\pm$ 1.32
	Top- k	88.83 $\pm$ 0.38	89.12 $\pm$ 0.56	88.11 $\pm$ 0.58	79.77 $\pm$ 0.61	80.38 $\pm$ 0.71	78.57 $\pm$ 0.81
	Eta	88.99 $\pm$ 0.30	88.48 $\pm$ 0.17	88.47 $\pm$ 0.26	80.15 $\pm$ 1.34	78.95 $\pm$ 0.63	79.39 $\pm$ 0.34
	Mirostat	89.24 $\pm$ 0.32	88.75 $\pm$ 0.39	89.09 $\pm$ 0.37	80.32 $\pm$ 0.86	80.12 $\pm$ 1.04	80.15 $\pm$ 0.26
	Typical	89.19 $\pm$ 0.12	88.85 $\pm$ 0.27	88.70 $\pm$ 0.24	80.06 $\pm$ 0.95	80.32 $\pm$ 0.63	79.45 $\pm$ 1.29
	Adaptive	88.98 $\pm$ 0.44	88.76 $\pm$ 0.33	88.83 $\pm$ 0.15	80.32 $\pm$ 0.75	80.79 $\pm$ 0.94	80.99 $\pm$ 0.60
	Min- p	89.39 $\pm$ 0.18	88.78 $\pm$ 0.56	88.40 $\pm$ 0.37	79.62 $\pm$ 0.97	80.55 $\pm$ 0.91	79.91 $\pm$ 0.59
	Ours	89.30 $\pm$ 0.46	88.85 $\pm$ 0.33	88.94 $\pm$ 0.40	80.35 $\pm$ 0.99	80.52 $\pm$ 0.65	80.23 $\pm$ 1.09
	CS	88.75 $\pm$ 0.47	88.39 $\pm$ 0.64	88.14 $\pm$ 0.27	79.59 $\pm$ 0.79	79.33 $\pm$ 0.61	79.30 $\pm$ 0.22
	CD	88.52 $\pm$ 0.47	88.52 $\pm$ 0.57	87.98 $\pm$ 0.40	79.77 $\pm$ 1.15	79.45 $\pm$ 0.68	79.33 $\pm$ 1.15

Table 12: Full per-cell accuracy (mean $\pm$ std, %) for Table 1. All values from 5 independent runs.

Method	w/o thinking			w/ thinking		
	$\tau=0.5$	$\tau=0.8$	$\tau=1.0$	$\tau=0.5$	$\tau=0.8$	$\tau=1.0$
Temperature	61.49 $\pm$ 0.67	61.83 $\pm$ 0.78	61.91 $\pm$ 0.38	66.87 $\pm$ 0.24	66.55 $\pm$ 0.19	66.16 $\pm$ 0.61
Top- p	62.64 $\pm$ 0.55	62.38 $\pm$ 0.26	61.61 $\pm$ 0.54	66.94 $\pm$ 0.53	66.64 $\pm$ 0.71	66.31 $\pm$ 0.72
Top- k	61.93 $\pm$ 0.25	62.11 $\pm$ 0.59	61.59 $\pm$ 0.85	66.50 $\pm$ 0.23	66.10 $\pm$ 0.39	66.26 $\pm$ 0.54
Eta	62.58 $\pm$ 0.41	62.30 $\pm$ 0.46	62.44 $\pm$ 0.58	66.72 $\pm$ 0.40	66.75 $\pm$ 0.74	66.70 $\pm$ 0.34
Mirostat	62.71 $\pm$ 0.22	62.52 $\pm$ 0.72	62.70 $\pm$ 0.54	66.61 $\pm$ 0.50	66.64 $\pm$ 0.29	66.55 $\pm$ 0.21
Typical	62.56 $\pm$ 0.29	62.00 $\pm$ 0.48	62.09 $\pm$ 0.70	66.96 $\pm$ 0.22	66.35 $\pm$ 0.55	66.54 $\pm$ 0.23
Adaptive	62.68 $\pm$ 0.83	61.93 $\pm$ 0.78	62.08 $\pm$ 0.49	66.49 $\pm$ 0.31	66.73 $\pm$ 0.39	66.22 $\pm$ 0.44
Min- p	62.15 $\pm$ 0.65	62.87 $\pm$ 0.41	62.05 $\pm$ 0.76	66.84 $\pm$ 0.46	66.78 $\pm$ 0.31	66.75 $\pm$ 0.30
Ours	62.85 $\pm$ 0.48	62.55 $\pm$ 0.69	62.18 $\pm$ 0.45	66.64 $\pm$ 0.29	66.76 $\pm$ 0.25	66.75 $\pm$ 0.45
CS	62.03 $\pm$ 0.59	62.00 $\pm$ 0.49	61.33 $\pm$ 0.50	66.61 $\pm$ 0.14	66.63 $\pm$ 0.58	66.16 $\pm$ 0.23
CD	60.58 $\pm$ 0.73	61.46 $\pm$ 0.54	62.02 $\pm$ 0.27	66.49 $\pm$ 0.43	66.64 $\pm$ 0.32	66.31 $\pm$ 0.59

Table 13: Full per-cell accuracy (mean $\pm$ std, %) for Table 2. All values from 5 independent runs.