GUARD: Glocal Uncertainty-Aware Robust Decoding for Effective and Efficient Open-Ended Text Generation

Yuanhao Ding[*1] **Esteban Garces Arias**[*2,3] **Meimingwei Li**[*2] **Julian Rodemann**[2,4] **Matthias Aßenmacher**[2,3] **Danlu Chen**[5] **Gaojuan Fan**[1] **Christian Heumann**[2] **Chongsheng Zhang**[†1]

[1]Henan University, [2]Department of Statistics, LMU Munich
[3]Munich Center for Machine Learning (MCML)
[4]CISPA Helmholtz Center for Information Security, Saarbrücken
[5]University of California, San Diego
{yhding, fangaojuan, cszhang}@henu.edu.cn
{esteban.garcesarias, matthias}@stat.uni-muenchen.de, M.Li@campus.lmu.de
{J.Rodemann, Christian.Heumann}@lmu.de, danlu@ucsd.edu

Abstract

Open-ended text generation faces a critical challenge: balancing coherence with diversity in LLM outputs. While contrastive search-based decoding strategies have emerged to address this trade-off, their practical utility is often limited by hyperparameter dependence and high computational costs. We introduce GUARD, a self-adaptive decoding method that effectively balances these competing objectives through a novel "Glocal" uncertainty-driven framework. GUARD combines global entropy estimates with local entropy deviations to integrate both long-term and short-term uncertainty signals. We demonstrate that our proposed global entropy formulation effectively mitigates abrupt variations in uncertainty, such as sudden overconfidence or high entropy spikes, and provides theoretical guarantees of unbiasedness and consistency. To reduce computational overhead, we incorporate a simple yet effective token-count-based penalty into GUARD. Experimental results demonstrate that GUARD achieves a good balance between text diversity and coherence, while exhibiting substantial improvements in generation speed. In a more nuanced comparison study across different dimensions of text quality, both human and LLM evaluators validated its remarkable performance. Our code is available at `https://github.com/YecanLee/GUARD`.

1 Introduction

Neural text generation models based on the Transformer decoder (Vaswani et al., 2017) have revolutionized natural language generation tasks such as story creation (Fang et al., 2021), context completion (Leng et al., 2024), or dialogue generation (Su et al., 2021). Decoding strategies play a crucial role in text generation with large language models (LLMs), affecting both the quality of the generated text and the computational efficiency during inference. The most commonly used strategies can be broadly categorized in deterministic (Freitag and Al-Onaizan, 2017; Carlsson et al., 2025) and stochastic approaches (Fan et al., 2018; Nguyen et al., 2025a; Aichberger et al., 2025). Deterministic approaches are likelihood-based techniques that typically (over-)emphasize coherence at the cost of diversity and are prone to producing repetitive text, a phenomenon referred to as *degeneration*. Stochastic approaches, on the other hand, aim to improve the diversity of texts, potentially sacrificing coherence and leading to semantic inconsistencies (Welleck et al., 2020). This inherent **coherence vs. diversity trade-off** (Garces Arias et al., 2025a) has motivated recent work to develop decoding strategies able to balance these two competing objectives. In particular, Contrastive Search (CS; Su et al., 2022) introduced a weighted combination of model confidence (favoring coherence) and a degeneration penalty (promoting diversity) controlled by fixed, tunable hyperparameters throughout the generation. Adaptive Contrastive Search (ACS; Garces Arias et al., 2024) addressed this hyperparameter dependence by dynamically adjusting them based on the local uncertainty of the model. In doing so, this approach still has its limitations as it relies only on local uncertainty information, and it substantially increases the computational overhead at inference time. In this study, we aim to address the following research questions (RQ):

1. **RQ1:** Can we improve text quality in open-ended text generation compared to existing approaches (§4.2)?

[*] Equal contribution
[†] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7202–7226
November 4-9, 2025

2. **RQ2:** Can we smooth out abrupt entropy deviations during generation without compromising statistical properties such as unbiasedness and consistency (§3.1)?

3. **RQ3:** Can we substantially reduce the latency and computational cost compared to (A)CS (§4.2, "Generation Speed")?

To address the above questions, we introduce a novel concept called "**Glocal**" uncertainty, which combines *global* entropy estimates with *local* entropy deviations to integrate long-term and short-term uncertainty signals, thereby maintaining robust decoding while ensuring statistical unbiasedness and consistency. Based on Glocal uncertainty, we propose a novel decoding strategy, **Glocal Uncertainty-Aware Robust Decoding (GUARD)**, which effectively balances the coherence and diversity of the generated texts (**RQ1**), addresses the smoothness of entropy estimates (**RQ2**), and improves inference speed (**RQ3**).

Our Contributions can be summarized as follows:

1. We introduce "Glocal" uncertainty, which integrates global entropy and instantaneous uncertainty signals to smooth out strong deviations in entropy, resulting in more robust decoding. We provide statistical guarantees (unbiasedness and consistency) for the proposed global entropy estimator, which is an integral part of the Glocal uncertainty.
2. Based on Glocal uncertainty, we design GUARD, a self-adaptive decoding strategy that effectively and efficiently balances coherence and diversity. Compared to (A)CS, it reduces the computational cost remarkably by incorporating a simple, yet effective, token count penalty.
3. We evaluate GUARD through comprehensive experimentation that spans diverse models, datasets, and decoding configurations, comparing it against state-of-the-art (SOTA) decoding strategies. Our assessment employs automatic, human, and LLM evaluations, which consistently show that GUARD represents a high-performing decoding strategy.

2 Related work

Deterministic Strategies primarily rely on maximizing the sequence probability. Greedy Search selects tokens based on pure maximum likelihood, while Beam search (Freitag and Al-Onaizan, 2017) improves upon this by maintaining multiple candidates simultaneously, ultimately yielding the most probable sequence. Further research (Vijayakumar et al., 2018; Holtzman et al., 2020; Wiher et al., 2022; Shi et al., 2024; Garces Arias et al., 2025b) has demonstrated that this often results in text degeneration and repetitive patterns. Recent developments, such as *hyperfitting* (Carlsson et al., 2025), show promising solutions to this problem, even within greedy search frameworks.

Stochastic Strategies based on sampling have emerged as a response to the limitations of deterministic approaches, particularly addressing text monotony and degeneration. Temperature sampling (Ackley et al., 1985) introduces a hyperparameter τ that modulates the sharpness of the output distribution, enabling control over generation diversity. Top-k (Fan et al., 2018) and Top-p sampling Holtzman et al. (2020) truncate the output distribution to avoid sampling from less reliable tail probabilities. Typical sampling (Meister et al., 2023) constrains the sampling distribution to tokens whose negative log-probabilities fall within a specific range of the model's conditional entropy, while min-p sampling (Nguyen et al., 2025a) introduces dynamic truncation based on the model's confidence. Adaptive Decoding (AD, Zhu et al., 2024) employs an entropy-based confidence metric to optimize candidate selection. While these sampling methods effectively reduce degeneration, they risk inconsistency, compromising overall text quality (Basu et al., 2021).

Contrastive Strategies typically aim at balancing coherence and diversity in text generation tasks. Contrastive Decoding (CD, Li et al., 2023) leverages the differential between expert and amateur language models to select tokens that maximize their log-likelihood difference. CS evaluates top-k tokens using a combination of model confidence and degeneration penalty to enhance output diversity. ACS builds upon CS by eliminating hyperparameters, addressing the unexplored impact of hyperparameter selection on generation quality (Garces Arias et al., 2025b), at the cost of additional computational overhead. However, its focus on immediate uncertainty overlooks temporal fluctuations that could benefit from more robustness. These limitations motivate GUARD, a novel strategy that enables an efficient, self-adaptive uncertainty-guided decoding.

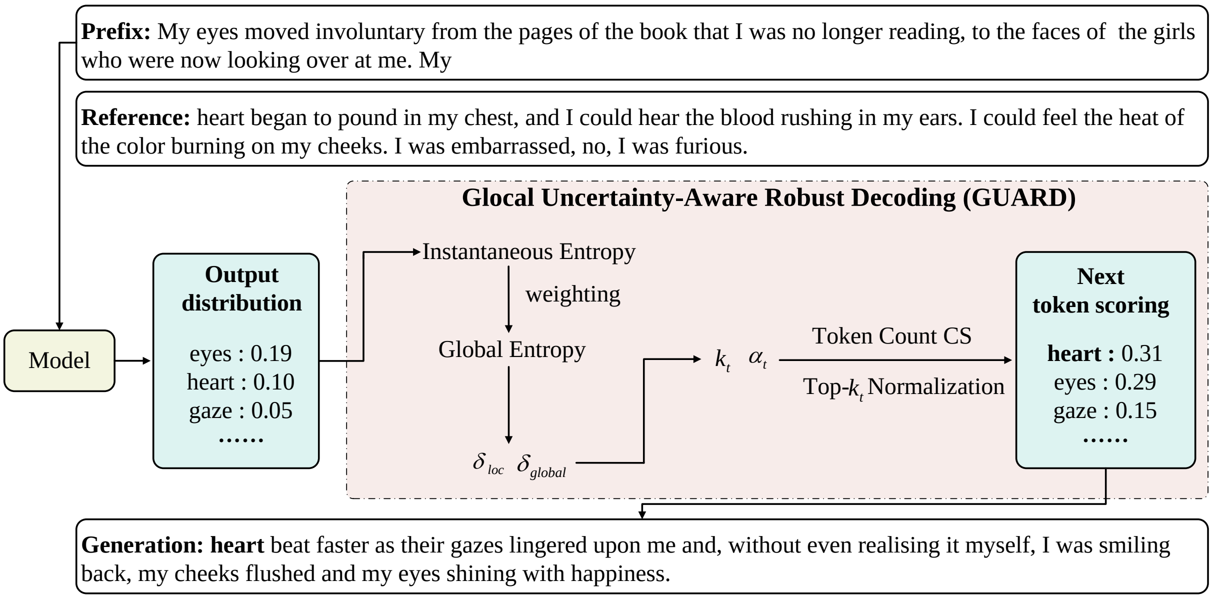


Figure 1: GUARD leverages local and global Shannon entropy deviations as proxies for model uncertainty, automatically adjusting CS parameters and incorporating a token count penalty for next-token selection.

3 Methodology

Figure 1 illustrates GUARD, which (a) leverages global entropy $H_{\text{glob},\,t}$ (Eq. (1)) to capture model uncertainty fluctuations throughout the decoding process and (b) introduces Glocal uncertainty (denominator in Eq. (6)). It thus considers both long- and short-term uncertainty fluctuations to enable adaptive control over k_t and α_t during (A)CS for the top-k_t most semantically accurate tokens. The method incorporates a simple yet effective token count-based penalty for repeated tokens.

3.1 Global Entropy and Statistical Properties

In the related ACS strategy, the Shannon entropy $H(X)_t = -\sum_{x\in\mathcal{V}} p(x \mid \boldsymbol{x}_{<t}) \ln p(x \mid \boldsymbol{x}_{<t})$ is used to dynamically adjust hyperparameters at the time step t. As is customary, X denotes a discrete random variable with $\mathcal{V}$ referring to the tokens; $t \in \{1, \dots, T\}$ denotes a single generation step, limited to the maximum sequence length (e.g., $T = 256$).

However, the Shannon entropy $H(X)_t$ only measures *local* uncertainty, i.e., the entropy at each generation time step $t \in \{1, \dots, T\}$. As illustrated in Figure 2 (left, solid line), local entropy measurements exhibit marked volatility. Addressing **RQ2**, we investigate whether these abrupt "spikes" can be effectively attenuated by a moving average of instantaneous entropy values, which we define as **global entropy** (dashed line, left part of Figure 2). We define global entropy $H_{\text{glob},\,t}$ at time step t in Eq. (1) as the weighted average of the Shannon entropies (measuring local uncertainty) throughout the generation process:

$$H_{\text{glob},t} = \frac{\sum_{i=1}^{t}(\lambda^{t-i} \cdot H(X)_i)}{\sum_{i=1}^{t}\lambda^{t-i}} \qquad (1)$$

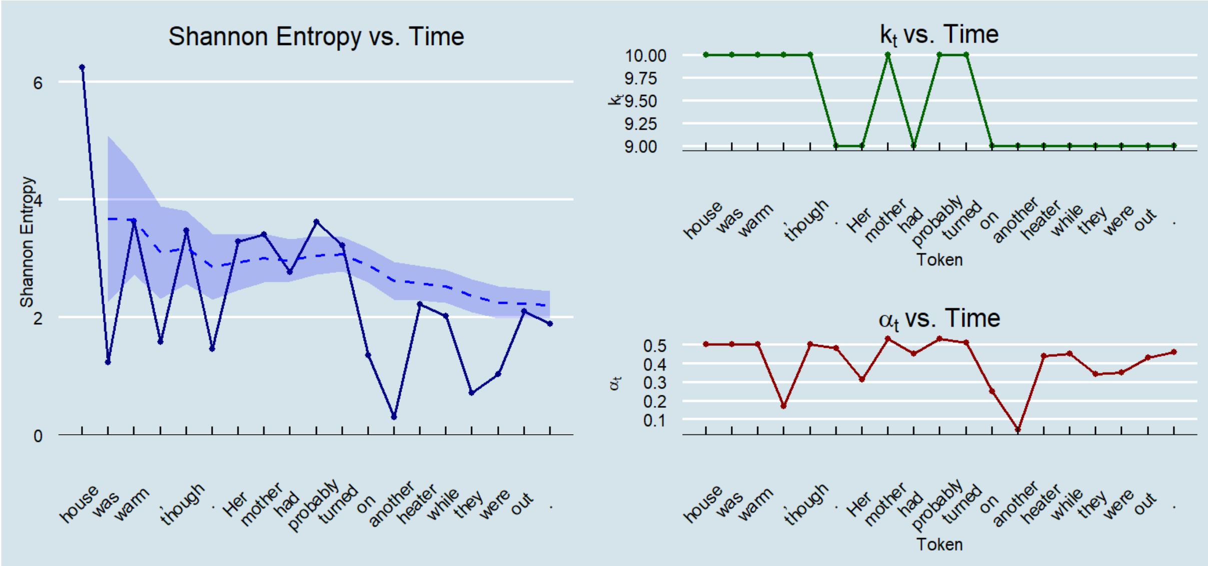
The parameter $\lambda \in (0, 1]$ denotes a weighting coefficient. We employ λ^{t-i} to reflect the temporal evolution of model uncertainty over time (i.e., over sequential generation steps), which can dynamically adjust the contribution of each generation step to $H_{\text{glob},t}$.

In the following, we will provide Proposition 1 to prove that $H_{\text{glob},t}$ is an unbiased estimator of the instantaneous entropy $H(X)_t$, irrespective of the concrete choice of $\lambda \in (0, 1]$. In Proposition 2, we further prove that $H_{\text{glob},t}$ is a consistent estimator of $H(X)_t$ under reasonable assumptions. This tells us that $H_{\text{glob},t}$ converges in probability, i.e., as $t \to \infty$, to the instantaneous entropy. Thus, it is not only unbiased but also has diminishing variance when the generation is sufficiently long. Detailed proofs of both results can be found in Appendix A.

Proposition 1 (Unbiasedness). Assume that the process $\{H(X)_t\}_{t\geq 1}$ is stationary. Then,

$$\mathbb{E}\big[H_{\text{glob},t}\big] = \mathbb{E}[H(X)_t] \qquad (2)$$

for any $\lambda \in [0, 1)$.



Prompt: *A shiver ran through her and she walked back inside. There was a chill in the air and she was only wearing a t-shirt and jeans. The*
Generated story: *house was warm, though. Her mother had probably turned on another heater while they were out.*

Figure 2: **Left:** Local (solid) and global (dashed/interval) Shannon entropy over the time steps of the generation. **Right:** Strong changes are smoothed to provide a robust entropy estimation before computing k and α over time, leading to increased stability.

Sketch of Proof. The result directly follows from the expectation's linearity and $\{H(X)_t\}_{t\geq 1}$ being stationary. Linearity gives

$$\mathbb{E}\big[H_{\text{glob},t}\big] = \sum_{i=1}^{t} \frac{\lambda^{t-i}}{\sum_{s=1}^{t} \lambda^{t-s}} \mathbb{E}\big[H(X)_i\big] \quad (3)$$

Stationarity implies

$$\forall t \in \{1, \dots, T\}: \ \mathbb{E}\big[H(X)_t\big] = H_0, \quad (4)$$

from which the claim obviously follows. □

Proposition 2 (Consistency). Assume the process $\{H(X)_t\}_{t\geq 1}$ is stationary, ergodic and bounded. Further assume that the covariances $\text{Cov}\big(H(X)_t, H(X)_s\big)$ decay sufficiently fast as $|t-s| \to \infty$. Then, as $t \to \infty$,

$$H_{\text{glob}} \xrightarrow{\mathbb{P}} \mathbb{E}[H(X)_t] \quad (5)$$

Sketch of Proof. We have that all $H(X)_t$ are bounded and their covariances decay sufficiently fast as $|t-s|$ increases. Moreover, the weights form a convex combination. Under standard mixing conditions, we show that

$$\text{Var}\big(H_{\text{glob}}\big) = \mathcal{O}\Big(\frac{1}{t}\Big),$$

where $\mathcal{O}$ is the Landau ("big O") notation. Then, by Chebyshev's inequality, for any $\epsilon > 0$ we have

$$\mathbb{P}\Big(\big|H_{\text{glob}} - \mathbb{E}[H_{\text{glob}}]\big| > \epsilon\Big) \leq \frac{\text{Var}\big(H_{\text{glob}}\big)}{\epsilon^2} \to 0$$

as $t \to \infty$, from which the claim follows. □.

Brief Summary. In Proposition 1 and Proposition 2, we prove that $H_{\text{glob},t}$ is not only unbiased, but also has diminishing variance when the generation is long enough. Such statistical properties support that we can reasonably estimate the instantaneous entropy with $H_{\text{glob},t}$. This answers **RQ2** – the intended smoothing of the uncertainty estimation does not come at the price of losing statistical validity. However, note that both results concern our internal estimators of the entropy, not the inferential properties of the final language model.

Interpretation. Our method employs $\lambda^{(t-i)}$ to reflect the temporal evolution of model uncertainty. Specifically, as the time step i approaches the current moment t, its contribution to H_{glob} increases. Conversely, time steps distant from t have diminishing effects on H_{glob}. The $\lambda^{(t-i)}$ factor implements a temporal decay mechanism, dynamically adjusting each time step's contribution to H_{glob}. We later set $\lambda = 0.95$, but recall that our method's unbiasedness and consistency hold for any $\lambda \in (0, 1]$.[1] Therefore, our Global entropy measure can effectively estimate local uncertainties in a statistically unbiased and consistent way.

[1] We experimentally verified that performance is relatively insensitive to the chosen value of λ (cf. Tab. 6, Appendix B) and thus set $\lambda = 0.95$ as an intermediate value.

3.2 Glocal Uncertainty and Candidate Token Set

To determine k_t, that is, the selection of candidate tokens in CS, we introduce *Glocal* uncertainty. It simultaneously considers both the *global* and *local* uncertainty fluctuations in the token generation steps, denoted as δ_{glob} and δ_{loc}, respectively. Specifically, we present Eq. (6) to automatically derive k_t for the Top-k_t candidate tokens.

$$k_t = 10 \cdot (1 - \frac{1}{\exp\left(\lambda_k \cdot \delta_{\text{loc}} + (1-\lambda_k) \cdot \delta_{\text{glob}}\right) + 1}) + 5 \tag{6}$$

$$\delta_{\text{loc}} = q \cdot \operatorname{arctanh}\left(\frac{H(X)_t - \operatorname{med}\left(H(X)_{t-w:t-1}\right)}{\ln|\mathcal{V}|}\right) \tag{7}$$

$$\delta_{\text{glob}} = q \cdot \operatorname{arctanh}\left(\frac{\operatorname{med}\left(H(X)_{t-w:t-1}\right) - \operatorname{med}\left(H_{\text{glob},t}\right)}{\ln|\mathcal{V}|}\right) \tag{8}$$

$$\lambda_k = \frac{|\delta_{\text{loc}}|}{|\delta_{\text{loc}}| + |\delta_{\text{glob}}| + \epsilon} \tag{9}$$

where the constants 10 and 5 for the parameter k_t define an effective range for exploration (Garces Arias et al., 2024). δ_{loc} measures the model's instantaneous uncertainty fluctuation compared to recent steps, i.e., the instantaneous uncertainty at the current t^{th} step $H(X)_t$ versus the median uncertainty in the recent time window w,[2] denoted as *med* $(H(X)_{t-w:t-1})$. δ_{glob} assesses the overall variation between the median of the short-term entropy *med* $(H(X)_{t-w:t-1})$ and the median of the long-term uncertainty trend up to the current t^{th} step *med* $(H_{\text{glob},t})$.[3] λ_k refers to the dynamic adjustment of local and global entropy during the generation process. q is an adaptive temperature parameter influencing the dynamic adjustment magnitude of k_t and α_t. It adaptively adjusts by integrating changes in both current and global uncertainties:

$$q = 1.0 + r_{\text{change}} + r_{\text{difference}} \tag{10}$$

r_{change} measures the rate of uncertainty change at t:

$$r_{\text{change}} = \frac{|H(X)_t - H(X)_{t-1}|}{H(X)_{t-1} + \epsilon} \tag{11}$$

$r_{\text{difference}}$ assesses the difference between local uncertainty within the current time window and the overall trend:

$$r_{\text{difference}} = \frac{|\text{med}(H(X)_{t-w:t-1}) - \text{med}(H_{\text{glob,t}})|}{\text{med}(H(X)_{t-w:t-1}) + \epsilon} \tag{12}$$

When r_{change} or $r_{\text{difference}}$ is large, q increases, causing δ_{loc} and δ_{glob} to be amplified. λ_k and λ_α dynamically allocate weights based on their ratio to adjust k_t and α_t, which are part of the token count penalty. When uncertainty is high, k_t increases, potentially introducing repetitive candidate words, but also α_t increases, amplifying the degradation penalty and counteracting the effect of a higher k_t. Conversely, a decrease of k_t and α_t results in outputs that favor high-confidence candidate words.

Rationale of Glocal Uncertainty. In Eq. (6), the denominator $C_{\text{sum}} = \exp(\lambda_k \cdot \delta_{\text{loc}} + (1-\lambda_k) \cdot \delta_{\text{glob}})$ can be regarded as an alternative to perplexity (Jelinek et al., 1977), since the latter is the exponential of the total uncertainty in the prediction up to the current t^{th} step, where Glocal uncertainty calculates the variations between the instantaneous entropy at the t^{th} step with respect to short-term median entropy, as well as the short- vs. longer-term entropy variations. Similar to perplexity, a lower C_{sum} value is preferred. When pronounced uncertainty variations occur, higher perplexity is indicated, suggesting we should select a larger k_t value to expand the candidate token set, accommodating model instability. Conversely, when variations are minimal, a smaller k_t value suffices, as the model demonstrates stability, allowing us to evaluate fewer candidate tokens.

Once k_t is determined, we select the adequate Top-k_t candidate tokens $\mathcal{V}^{(k_t)}$, then design a token count penalty strategy to increase the diversity of the generated texts by reducing repetitiveness. Rather than employing a cosine-similarity-based penalty as in (A)CS, GUARD (as defined in Eq. (13)) circumvents the over-penalization of semantic similarity, leads to a more effective reduction of repetitions (and hence increased diversity), and accelerates the generation speed:

$$P(x_{next}) = \underset{v \in \mathcal{V}^{(k_t)}}{\arg\max} \left\{ p_\theta(v|\boldsymbol{x}_{<k_t}) \times \underbrace{\alpha_t^{token\ counts(v)}}_{\text{degeneration penalty}} \right\} \tag{13}$$

where α_t is the penalty coefficient and token counts(v) denotes the frequency of the candidate v in the previously generated sequence. To

[2] To select the optimal value for w, we conducted extensive experiments, which are presented in Appendix B (Tables 8 – 14). We observe that w does not notably influence text quality. Ultimately, we set $w = 7$, keeping it constant for our experiments.

[3] The choice of *median* and *arctanh* was experimentally validated. Our experiments compared different pairings, including *winsorized mean*, *mean*, *median*, *logarithmic map*, and *arctanh*, with detailed results shown in Table 7, Appendix B.

dynamically determine α_t, we reuse the core of Eq. (6), replacing $|\mathcal{V}|$ with k_t in Eq. (7) and (8):

$$\alpha_t = 1 - \frac{1}{\exp\left(\lambda_k \cdot \delta_{\text{loc}} + (1 - \lambda_k) \cdot \delta_{\text{glob}}\right) + 1} \tag{14}$$

When $\mathcal{V}^{(k_t)}$ is small, each repetitive token in $\mathcal{V}^{(k_t)}$ will receive a very light penalty, and vice versa. Appendix C illustrates the entire algorithm as pseudo-code to facilitate understanding of its flow. In Figure 2, we demonstrate that fluctuations can be smoothed to ensure a robust entropy estimation (left) when using our technique in determining k_t and α_t over time (right). Our experiments show that the token count-based penalty strategy can substantially reduce repetitions, therefore increasing the diversity of the text outputs while enhancing text generation speed compared to using cosine similarity.

Self-Adaptive Design of GUARD. GUARD achieves self-adaptivity for all quality-determining hyperparameters in CS methods; while it exhibits hyperparameters λ and w, extensive testing (Appendix B) confirms that specific choices have negligible impact on generation quality, unlike other decoding methods (Garces Arias et al., 2025b).

4 Experiments

4.1 Experimental Setup

This section will introduce the metrics and datasets used in our experiments, as well as the baseline decoding strategies we use for reference and the LLMs we employ.

LLMs. We employ six LLMs: Qwen2.5-7B (Qwen, 2024), Deepseek-llm-7B-base (DeepSeek-AI et al., 2025), Mistral-v0.3 (Jiang et al., 2023), Llama-3 (Grattafiori et al., 2024), Llama-2 (Touvron et al., 2023), Gemma-7B (Team et al., 2024).

Evaluation Metrics. We employ the following five commonly used metrics to automatically assess generation quality: Diversity (Su et al., 2022), MAUVE (Pillutla et al., 2021), Coherence (Su et al., 2022), BERTScore (Zhang et al., 2019). Formulas are provided in Appendix D.

Human Evaluation. To evaluate the quality of the generated text, human evaluators assessed two key aspects: semantic coherence and fluency. Four native English speakers evaluated 600 pairs of competing text continuations, which were uniformly distributed across the three datasets. We measured inter-rater agreement using Fleiss' Kappa and performed exact binomial tests to determine statistically significant differences between ACS and GUARD. Further details on the evaluation are provided in Appendix E.

LLM Evaluation. We employ LLM-as-a-judge (judge: GPT-4) to systematically evaluate the quality of the generated text on $n = 150$ continuations, uniformly distributed across the three datasets. This method aims to approximate the evaluation process of human experts on a large scale. For this, we measured six core metrics: fluency, coherence, factuality, informativeness, interestingness, and story development. The prompt is provided in Appendix F.

Datasets and Baselines. We focus on three datasets for open-ended text generation from different domains: Wikinews (*n = 2000*), Wikitext (*n = 1314;* Merity et al., 2016), and BookCorpus (*n = 1947;* Zhu at al., 2015). We compare GUARD to SOTA LLM decoding methods from three categories: Deterministic (greedy search, beam search with $B \in \{3, 5, 10, 15, 20, 50\}$), stochastic (temperature sampling with $\tau = 0.9$, top-k sampling with $k \in \{3, 5, 10, 15, 20, 50\}$, top-$p$ sampling with $p \in \{0.60, 0.70, 0.80, 0.90, 0.95\}$, typical sampling with $\tau = 0.2$, adaptive decoding), and contrastive approaches (CD, CS with $k \in \{5, 10\}$, $\alpha = 0.6$, and ACS with $q = 1$). Across all experiments, we used a maximum token length of 256 tokens for the text generations.

Hardware. We used NVIDIA 4090 (24GB memory) and NVIDIA H100 (80GB memory).

4.2 Results

Automatic evaluation results. The automatic evaluations of different methods, based on diversity, MAUVE, and coherence, are presented in Table 1. We observe that, in general, GUARD outperforms all other competing strategies in both diversity and coherence. In terms of MAUVE, it is roughly on par with the best competing methods on each dataset. More importantly, it achieves results closest to human references without relying on specific hyperparameter choices. We observe that GUARD generally performs well according to BERTScore, which, however, is not very discriminative. Furthermore, we further validate the robustness of our method for other model architectures (cf. Appendix

Strategy	Wikitext (*n* = *1314*)				Wikinews (*n* = *2000*)				BookCorpus (*n* = *1947*)			
	Div.	MAUVE	Coh.	BERT-Sc.	Div.	MAUVE	Coh.	BERT-Sc.	Div.	MAUVE	Coh.	BERT-Sc.
Greedy	16.97	48.28	−0.78	81.45	28.13	74.91	−0.90	82.51	4.07	17.99	−0.65	79.92
Beam (5)	14.42	26.39	−0.77	81.20	24.94	54.31	−0.87	83.23	5.39	18.69	−0.56	80.03
Temp. (0.9)	84.81	92.85	**-2.43**	81.86	92.71	**96.29**	−2.41	83.16	89.58	85.54	−2.33	81.78
Top-k (50)	82.69	88.97	−1.98	80.64	92.13	94.52	−2.01	83.17	91.54	93.24	−2.52	81.29
Top-p (0.95)	82.47	86.37	−2.11	81.50	91.47	96.09	−2.15	82.84	88.92	**94.81**	−2.29	81.31
Typical	69.54	81.57	−1.78	81.34	82.31	95.66	−1.83	**83.86**	96.29	88.58	−3.68	80.06
CD	72.70	79.76	−2.61	81.72	75.28	76.20	−2.36	82.57	76.00	72.10	**-2.53**	81.51
AD	87.19	**92.20**	−2.51	81.83	91.49	94.38	−2.35	82.52	90.70	94.49	−2.49	81.82
CS $(5, 0.6)$	72.08	80.52	−1.43	81.53	87.47	91.47	−1.31	83.01	68.26	87.92	−1.24	81.20
CS $(10, 0.6)$	78.68	73.50	−1.69	80.57	91.38	90.49	−1.57	83.09	82.60	75.53	−1.57	81.22
ACS ($q = 1$)	81.57	78.82	−1.95	80.00	**92.79**	91.23	−1.72	81.50	87.86	78.32	−1.52	79.79
GUARD	**92.86**	90.82	−2.61	**81.90**	95.20	93.60	**-2.38**	83.16	**96.18**	92.59	**-2.53**	**81.83**
Human	93.84	100.00	−2.43	100.00	93.50	100.00	−2.86	100.00	94.98	100.00	−2.99	100.00

Table 1: Averaged automatic evaluation results for Qwen2.5-7B. Hyperparameters for competing strategies chosen based on experimental evaluation (Appendix G, Tables 18, 19, 20) or based on the original paper (CS, ACS). Scores closest to human highlighted in **bold**.

Dataset (*n* = *200* each)	Human Evaluation			
	Sem. Coh.* (%)↑		Fluency** (%)↑	
	ACS	GUARD	ACS	GUARD
Wikitext	37.50	**62.50**	43.75	**56.25**
Wikinews	40.00	**60.00**	45.25	**54.75**
BookCorpus	48.75	**51.25**	**52.50**	47.50
All	42.00	**58.00**	47.00	**53.00**

Table 2: **Human evaluation:** Share of human evaluators favoring a strategy w.r.t. perceived semantic coherence and fluency. P-values for the exact binomial test: $*p_{\text{Sem. Coh.}} = 2.875e-06$, $**p_{\text{Fluency}} = 0.01819$. Best results are highlighted in **bold**.

G, Tables 16 – 17). Notably, GUARD is well-suited for decreasing the repetitive use of single tokens, as indicated by the high diversity values across all three benchmark datasets. While one might suspect that this improvement comes at the cost of coherence, our experiments show that GUARD also consistently exhibits coherence values that are very close to those of the human-written completions, suggesting a stronger calibration towards human production (Garces Arias et al., 2025c).

Human evaluation. We only compare ACS and GUARD for the sake of practicability and since both show strong performance (cf. Table 1. The results of this study are summarized in Table 2: Overall, GUARD demonstrates superior performance compared to ACS in terms of both semantic coherence and fluency, with the sole exception of fluency on the BookCorpus dataset. Specifically, human preferences indicate that GUARD outperforms ACS on average by approximately 6% in semantic coherence and 16% in fluency across all datasets. Both differences are statistically significant at the $\alpha = 0.05$ level. The inter-rater reliability (Fleiss' κ = 0.41) shows moderate agreement among evaluators. In Appendix H, we include a small case study for the interested reader that further substantiates these findings. Overall, the human evaluation comprehensively addresses **RQ1**, confirming that both automatic metrics and human judgments favor our approach over alternative methods.

LLM evaluation. We include LLM evaluation as an approximation of human judgments, comparing GUARD to its natural baseline ACS, as well as Top-k and Top-p. The evaluation focuses on the six dimensions described in the experimental setup. The results (cf. Table 3) demonstrate that GUARD outperforms all three competitors in five out of six metrics (except factuality), where ACS has a slight edge, and comparison to the other two strategies results in a tie. When comparing human to LLM judgments for coherence and fluency (cf. Table 4), results (a) demonstrate strong agreement and (b) show that human evaluators indicate an even stronger preference for GUARD's coherence compared to LLM-judges.

Generation speed. In Table 5, we compare GUARD to CS and ACS in terms of the average generation speed. Based on 30 generation samples, all three methods produce generations of comparable lengths. CS spends 11.6 seconds per story, ACS takes an even longer, from 15.7 seconds to 16.3 seconds (across different temperature settings),

Metric	GUARD wins	ACS wins	Tie	GUARD wins	Top-k wins	Tie	GUARD wins	Top-p wins	Tie
Overall	**34.3**	30.6	35.1	**65.8**	17.5	16.7	**49.2**	34.2	16.7
Fluency	**40.7**	35.3	24.0	**76.7**	23.3	0.0	**56.7**	43.3	0.0
Sem. Coh.	**40.7**	34.0	25.3	**80.0**	20.0	0.0	**58.3**	41.7	0.0
Factuality	4.0	**10.0**	86.0	0.0	0.0	**100.0**	0.0	0.0	**100.0**
Informativeness	**40.0**	35.3	24.7	**78.3**	21.7	0.0	**60.0**	40.0	0.0
Interestingness	**40.0**	34.7	25.3	**80.0**	20.0	0.0	**60.0**	40.0	0.0
Story Development	**40.7**	34.0	25.3	**80.0**	20.0	0.0	**60.0**	40.0	0.0

Table 3: Win shares (%) of GUARD vs. ACS/Top-k/Top-p for LLM evaluation for $n = 150$ continuations. Best results are highlighted in **bold**.

Metric	Dataset ($n = 50$ each)	Human Evaluation			LLM Evaluation			Agree?
		GUARD (%)	ACS (%)	Ties (%)	GUARD (%)	ACS (%)	Ties (%)	
Sem. Coh.	Wikitext	**44.0**	19.0	37.0	**42.0**	32.0	26.0	✓
	Wikinews	**43.5**	23.5	33.0	**40.0**	34.0	26.0	✓
	BookCorpus	**38.0**	35.5	26.5	**40.0**	36.0	24.0	✓
	Overall ($n = 150$)	**41.8**	26.0	32.2	**40.7**	34.0	25.3	✓
Fluency	Wikitext	**23.0**	10.5	66.5	**42.0**	32.0	26.0	✓
	Wikinews	**21.5**	12.0	66.5	**40.0**	**40.0**	20.0	✗
	BookCorpus	13.0	**18.0**	69.0	**40.0**	34.0	26.0	✗
	Overall ($n = 150$)	**19.2**	13.5	67.3	**40.7**	35.3	24.0	✓

Table 4: Win shares (Human and LLM evaluation) w.r.t. coherence and fluency of GUARD and ACS. Best results are highlighted in **bold**.

Method	sec/story	#tokens/sec	#tokens/story
CS (10, 0.6)	11.6	22.0	249.7
ACS (q = 1)	15.7	16.3	255.7
ACS (q = 2)	15.9	16.1	255.8
ACS (q = 8)	16.3	15.3	255.8
GUARD	**4.42**	**28.3**	**255.9**

Table 5: Average generation speed ($n = 30$) of GUARD with CS and ACS ($q \in \{1, 2, 8\}$). Best results in **bold**.

while GUARD only needs 4.42 seconds per story. We measure the average number of tokens decoded per second, and observe that CS decodes on average 21.98 tokens/second, ACS between 15.35 and 16.29, and GUARD decodes 28.3 tokens/second on average. This answers **RQ3**, since the generation speed of CS-based approaches can be substantially increased using GUARD.

5 Discussion

Stationarity Assumption. The strict stationarity assumption might be a strong idealization in real, dynamic text generation processes. However, the main purpose of introducing the stationarity assumption here is to provide a theoretical guarantee of unbiasedness for our $H_{glob,t}$ estimator (Proposition 1). This theoretical result aims to show that, under ideal conditions, our weighted average design does not introduce systematic bias. It is worth noting that the core advantage of our work does not entirely rely on this strict assumption. The key role of our design (i.e., the exponentially weighted moving average in Eq. (1)) in practice is to smooth out drastic fluctuations in entropy. The presence of the decay factor λ makes it more sensitive to recent entropy values, while gradually "forgetting" older ones. This mechanism allows $H_{glob,t}$ to dynamically adapt to local changes in the entropy process, providing a locally stationary and robust long-term trend estimate, even if the overall process is non-stationary. As shown in Figure 2, $H_{glob,t}$ effectively smooths out local entropy, illustrating that our method remains effective even in the presence of entropy fluctuations (non-stationary conditions).

Self-Adaptiveness. While we rightfully advertise our method's ability to dynamically adjust core decoding hyperparameters (k, α, and q) without manual tuning, GUARD still relies on λ and w. Nevertheless, our experiments suggest that neither of them substantially influences performance, and can be kept at their default values.

6 Conclusion

We introduce GUARD, a self-adaptive decoding strategy that adaptively balances text coherence and diversity by responding to fluctuations in long-term and short-term uncertainty. By integrating a token frequency penalty into CS, GUARD reduces repetitive outputs and computational overhead while maintaining context fidelity. Theoretical analysis confirms that our estimation of global entropy preserves key statistical properties, such as unbiasedness and consistency. Comprehensive experiments across multiple open-source models, datasets, metrics, human and LLM evaluations demonstrate improvements in coherence, fluency, diversity, efficiency, informativeness, interestingness, and story development over established decoding methods. Overall, GUARD offers a robust approach for improving open-ended text generation. Future work will extend our method to additional text-generation tasks. We believe that investigating its integration with supervised fine-tuned models and exploring its performance in multilingual and low-resource settings might yield valuable insights.

Limitations

While our proposed method demonstrates improvements in open-ended text generation quality, several limitations warrant acknowledgment:

(1) Our evaluation focuses exclusively on open-ended text generation tasks. The transferability of our approach to other NLP applications, such as summarization, machine translation, and classification, remains to be investigated.

(2) The empirical evaluation is confined to English-language datasets, leaving questions about cross-lingual generalizability, particularly for low-resource languages, unanswered.

(3) While we demonstrate effectiveness across various open-source models, we have not evaluated our method on proprietary language models, which may exhibit different behaviors.

(4) Our experiments are limited to base models. The impact of our approach on supervised fine-tuned models represents an important direction for future research.

(5) Finally, our setting was limited by a maximum length of 256 tokens, and the behaviour in long-text generation scenarios is yet to be explored.

Ethics Statement

We affirm that our research adheres to the ACL Ethics Policy. This work involves the use of publicly available datasets and does not include any personally identifiable information. For our human evaluation, we employed third-party evaluators, ensuring a rate of over $20 per hour. An ethical concern worth mentioning is the use of language models for text generation, which may produce harmful content, either through intentional misuse by users or unintentionally due to the training data or algorithms. We declare that there are no conflicts of interest that could potentially influence the outcomes, interpretations, or conclusions of this research. All funding sources supporting this study are acknowledged in the acknowledgments section. We have diligently documented our methodology, experiments, and results, and we commit to sharing our code, data, and other relevant resources to enhance reproducibility and further advancements in the field.

Acknowledgments

This work was partially supported by the MOE Liberal Arts and Social Sciences Foundation (No.23YJAZH210), Major Program of National Social Science Foundation (No.23&ZD309), Henan Provincial Center for Outstanding Overseas Scientists (No.GZS2025004), High Level Talent International Training Program of Henan Province (No.GCC2025010), Henan Provincial Science and Technology Project (No.252102210150), Key Scientific Research Project for Universities in Henan Province (No.25A520014), and the Chinese Scholarship Council (Grant No.202308410339). Moreover, Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI, under Grant No.460037581. Julian Rodemann acknowledges the funding support from the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics", the Bavarian Institute for Digital Transformation (bidt) and the Bavarian Academy of Sciences (BAdW) within a graduate scholarship. We also thank the support from the Munich Center for Machine Learning (MCML), and the Department of Statistics, LMU Munich. Last but not least, we thank all the anonymous reviewers and area chair(s) for their constructive feedback throughout the review process.

References

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

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2025. Improving uncertainty estimation through semantically diverse language generation.

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

NH Bingham. 1973. Independent and stationary sequences of random variables.

Fredrik Carlsson, Fangyu Liu, Daniel Ward, Murathan Kurfali, and Joakim Nivre. 2025. The hyperfitting phenomenon: Sharpening and stabilizing llms for open-ended text generation. In *2025 International Conference on Learning Representations (ICLR 2025)*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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, Melbourne, Australia. Association for Computational Linguistics.

Le Fang, Tao Zeng, Chao-Ning Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *ArXiv*, abs/2101.00828.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.

Esteban Garces Arias, Hannah Blocher, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025a. Towards better open-ended text generation: A multicriteria evaluation framework. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 631–654, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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

Esteban Garces Arias, Julian Rodemann, and Christian Heumann. 2025c. The geometry of creative variability: How credal sets expose calibration gaps in language models. In *Proceedings of the Second Workshop on Uncertainty-Aware NLP*, Suzhou, China. Association for Computational Linguistics.

Esteban Garces Arias, Julian Rodemann, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2024. Adaptive contrastive search: Uncertainty-guided decoding for open-ended text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15060–15080, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

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

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Siyi Leng, Zhenxin Zhang, and Liqiang Zhang. 2024. A point contextual transformer network for point cloud completion. *Expert Systems with Applications*, 24.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori 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, Toronto, Canada. Association for Computational Linguistics.

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*, abs/1609.07843.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025a. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *2025 International Conference on Learning Representations (ICLR 2025)*.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025b. Turning up the heat: Min-p sampling for creative and coherent llm outputs. *Preprint*, arXiv:2407.01082.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Qwen. 2024. Qwen2.5: A party of foundation models.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

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, Miami, Florida, USA. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Annual Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Yixuan Su, Yan Wang, Deng Cai, Simon Baker, Anna Korhonen, , and Nigel Collier. 2021. Prototype-to-style: dialogue generation with style-aware editing on retrieval memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. 2020. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5553–5568, Online. Association for Computational Linguistics.

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

Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. 2024. Privacy-preserving instructions for aligning large language models. *ArXiv*, abs/2402.13659.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

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*.

Zhu at al. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV 2015)*, pages 19–27.

Appendix

A Proofs

For ease of exposition, we restate all claims before proving them.

A.1 Proof of Proposition 1

Proposition (Unbiasedness of H_{glob}) Assume that the process $\{H(X)_t\}_{t\geq 1}$ is stationary. Then,

$$\mathbb{E}\big[H_{\text{glob}}\big] = \mathbb{E}[H(X)_t]$$

for any $\lambda \in [0, 1)$.

Proof 1 *Define the normalized weights*

$$w_i = \frac{\lambda^{(t-i)}}{\sum_{s=1}^{t} \lambda^{(t-s)}}, \quad \textit{such that} \quad \sum_{i=1}^{t} w_i = 1,$$

for any $\lambda \in [0, 1)$. Then we can write the estimator (1) as

$$H_{glob} = \sum_{i=1}^{t} w_i \, H(X)_i.$$

Taking expectation and using linearity together with the stationarity assumption, we obtain

$$\mathbb{E}\big[H_{glob}\big] = \sum_{i=1}^{t} w_i \, \mathbb{E}\big[H(X)_i\big] = \sum_{i=1}^{t} w_i \, H_0 = H_0.$$

Since under stationarity, the instantaneous entropy's expectation at t is $\mathbb{E}[H(X)_t] = H_0$, the claim immediately follows:

$$E\big[H_{glob}\big] = \mathbb{E}[H(X)_t].$$

□

A.2 Proof of Proposition 2

Proof 2 *We show that the variance of H_{glob} vanishes as $t \to \infty$, so that by Chebyshev's inequality the estimator converges in probability to $\mathbb{E}[H(X)_t]$ under the following assumptions*

(1) Stationarity and Ergodicity: *The sequence $\{H(X)_t\}_{t\geq 1}$ is stationary and ergodic with*

$$\forall t \in \{1, \ldots, T\}: \; E\Big[H(X)_t\Big] =: H_0.$$

(2) Boundedness: *There exists a constant $M < \infty$ such that*

$$|H(X)_t| \leq M \quad \textit{for all } t.$$

(3) Decaying Covariances: *The covariances $\text{Cov}\big(H(X)_t, H(X)_s\big)$ decay sufficiently fast as $|t-s|$ increases.*

Decaying covariances can be ensured by, e.g., strong mixing of the sequence $\{H(X)_t\}$ with mixing coefficients $\alpha(n)$ satisfying

$$\sum_{n=1}^{\infty} \alpha(n)^{\delta/(2+\delta)} < \infty$$

for some $\delta > 0$.

The argument is as follows. Define the normalized weights

$$w_i = \frac{\lambda^{t-i}}{\sum_{s=1}^{t} \lambda^{t-s}}, \quad \textit{such that} \quad \sum_{i=1}^{t} w_i = 1,$$

for any $\lambda \in [0, 1)$.
Since

$$H_{glob,t} = \sum_{i=1}^{t} w_i \, H(X)_t,$$

its variance is given by

$$\mathrm{Var}\big(H_{glob,t}\big) = \sum_{i=1}^{t} \sum_{s=1}^{t} w_i \, w_s \; \mathrm{Cov}\Big(H(X)_i, H(X)_s\Big).$$

Per assumptions, we have that all $H(X)_t$ *are bounded by* M *and their covariances decay sufficiently fast as* $|t-s|$ *increases. Moreover, the weights form a convex combination. Under standard mixing conditions (see, e.g., Bingham (1973)), one can show that*

$$\mathrm{Var}\big(H_{glob,t}\big) = \mathcal{O}\Big(\frac{1}{t}\Big),$$

where $\mathcal{O}$ *is the Landau ("big O") notation. Then, by Chebyshev's inequality, for any* $\epsilon > 0$ *we have*

$$\mathbb{P}\Big(\big|H_{glob} - \mathbb{E}[H_{glob}]\big| > \epsilon\Big) \leq \frac{\mathrm{Var}\big(H_{glob}\big)}{\epsilon^2} \to 0$$

where as $t \to \infty$.
Together with Proposition 1, this implies

$$H_{glob,t} \xrightarrow{\mathbb{P}} \mathbb{E}[H(X)_t]$$

as $t \to \infty$. *Thus,* $H_{glob,t}$ *is a consistent estimator.*

B Further (Design Choice) Experiments

Optimal λ Selection. We further investigate how different values of λ influence the quality of text generations. We find that model performance remains stable for $\lambda \in \{0.91, 0.93, 0.95, 0.97, 0.99\}$, thus we select $\lambda = 0.95$ as our default parameter. This experiment is based on the GPT-2 XL model (Radford et al., 2019) to balance computational expenses.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
$\lambda = 0.91$	95.25	83.64	-2.60	95.86	89.39	-2.34	96.41	83.43	-2.61
$\lambda = 0.93$	95.22	81.39	-2.60	96.31	92.79	-2.34	96.39	81.55	-2.61
$\lambda = 0.95$	95.64	85.55	-2.62	96.19	90.61	-2.35	96.35	85.83	-2.62
$\lambda = 0.97$	96.00	79.27	-2.63	96.28	92.32	-2.36	96.60	82.79	-2.63
$\lambda = 0.99$	95.95	80.48	-2.67	96.46	90.35	-2.39	96.75	79.73	-2.65

Table 6: Evaluation of performance across three datasets for λ values in (0.91, 0.99).

Median Aggregation and Arctanh. To validate our design choices of Median aggregation and Arctanh mapping, we conducted statistical significance testing across multiple aggregation and mapping combinations. Bootstrap analysis with 95% confidence intervals demonstrated that our chosen Median + Arctanh method achieves significantly superior MAUVE scores (93.60 [92.71, 94.48]) compared to all alternative approaches ($p < 0.001, t = 4.98$). The effect sizes for these MAUVE improvements are consistently large (all Cohen's $d > 1.75$), with percentage improvements ranging from 3.61% to 6.39% over alternatives. While other combinations achieve better Diversity scores and certain alternatives show marginal improvements in Coherence, we follow common practices to guide these choices (Hewitt et al., 2022; Zhu et al., 2024; Nguyen et al., 2025b; Yu et al., 2024; Meister et al., 2023). The Friedman test across all metrics ($\chi^2 = 9.10$) indicates no statistically significant differences in overall performance across methods, suggesting that trade-offs exist between optimization targets.

Aggregation	Mapping	Diversity	MAUVE	Coherence	Notes
Winsorized mean	Logarithmic map	95.22	88.44	-2.42	-
Median	Logarithmic map	95.71	90.16	-2.48	-
Winsorized mean	Arctanh	94.49	87.98	-2.35	-
Mean	Logarithmic map	95.75	88.91	-2.35	-
Mean	Arctanh	97.15	90.34	-2.61	-
Median	Arctanh	95.20	93.60	-2.38	GUARD (Ours)

Table 7: Evaluation of different value assignment strategies and function mappings using the Wikinews dataset.

Optimal Locality Window Selection. We evaluated the performance of various locality window sizes, $w \in \{2, 3, 4, 5, 6, 7, 8, 9\}$, with the results summarized in Table 8. The analysis indicates that a window size of $w = 7$ yields good performance, while other choices in that range also display similar scores. To further validate this finding, we conducted additional experiments using multiple models, including Qwen2.5-7B, Llama2, Deepseek-llm-7B-base, Llama-3, Mistral-v0.3, Gemma-7B, and GPT2-xl. Detailed results for these experiments are provided in Tables 8, 9, 10, 11, 12, 13, and 14.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh. (%)
GUARD (w=2)	93.56	87.89	-2.57	94.64	94.20	-2.35	95.94	91.25	-2.52
GUARD (w=3)	94.06	88.87	-2.67	96.17	94.29	-2.41	96.89	90.43	-2.61
GUARD (w=4)	93.11	89.32	-2.57	94.73	94.83	-2.32	95.55	93.58	-2.46
GUARD (w=5)	92.97	87.26	-2.61	96.05	93.56	-2.39	96.51	92.30	-2.54
GUARD (w=6)	92.92	86.85	-2.57	96.05	93.56	-2.34	95.69	93.57	-2.45
GUARD (w=7)	92.86	90.82	-2.61	95.20	93.60	-2.38	96.18	92.59	-2.52
GUARD (w=8)	92.51	88.29	-2.57	94.45	93.66	-2.32	95.48	91.56	-2.46
GUARD (w=9)	92.82	89.78	-2.61	94.94	94.53	-2.38	96.14	89.96	-2.52

Table 8: Qwen2.5-7B: The performance under different datasets and w values shows that w = 7 is an appropriate locality window. As the value of w increases, the data value decreases; hence, no further experiments are conducted.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
GUARD (w=2)	95.49	90.59	-2.67	96.19	93.21	-2.25	96.21	93.08	-2.62
GUARD (w=3)	96.73	91.46	-2.87	97.18	95.33	-2.36	97.10	92.76	-2.61
GUARD (w=4)	94.73	91.27	-2.64	96.20	92.90	-2.26	96.89	91.97	-2.64
GUARD (w=5)	95.82	87.91	-2.77	96.91	92.85	-2.33	95.59	92.46	-2.59
GUARD (w=6)	94.50	90.00	-2.66	96.45	93.41	-2.24	96.15	92.78	-2.71
GUARD (w=7)	95.60	89.52	-2.75	96.76	93.34	-2.31	96.21	93.08	-2.62
GUARD (w=8)	94.74	87.78	-2.66	96.48	90.41	-2.25	96.10	92.78	-2.49
GUARD (w=9)	95.38	90.78	-2.74	96.54	94.12	-2.31	95.98	92.12	-2.59

Table 9: Llama-2: The performance under different datasets and w values shows that w = 7 is the appropriate locality window.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
GUARD (w=2)	95.62	89.45	-2.65	96.56	96.50	-2.21	96.32	90.36	-2.58
GUARD (w=3)	97.21	88.12	-2.74	97.55	93.36	-2.33	97.34	91.93	-2.68
GUARD (w=4)	95.41	84.49	-2.54	96.64	94.31	-2.20	96.09	90.43	-2.50
GUARD (w=5)	96.82	89.74	-2.68	97.24	94.32	-2.29	97.02	89.17	-2.61
GUARD (w=6)	95.64	83.35	-2.48	96.59	93.77	-2.20	96.34	89.46	-2.52
GUARD (w=7)	96.55	88.96	-2.65	97.14	94.84	-2.23	96.86	90.07	-2.61
GUARD (w=8)	95.70	91.09	-2.56	96.71	94.27	-2.21	96.23	88.95	-2.52
GUARD (w=9)	96.52	85.69	-2.65	97.08	94.24	-2.27	96.90	91.80	-2.61

Table 10: Deepseek-llm-7B-base: The performance under different datasets and w values shows that w = 7 is the appropriate locality window.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
GUARD (w=2)	95.47	89.85	-2.57	96.16	94.88	-2.15	96.02	91.20	-2.49
GUARD (w=3)	96.65	89.74	-2.66	97.25	95.49	-2.24	96.84	91.52	-2.57
GUARD (w=4)	94.94	89.98	-2.45	95.82	96.12	-2.12	95.75	90.13	-2.41
GUARD (w=5)	95.95	88.72	-2.56	96.67	93.44	-2.21	96.39	92.50	-2.52
GUARD (w=6)	94.88	90.74	-2.47	95.58	94.51	-2.14	95.73	91.14	-2.44
GUARD (w=7)	95.84	91.25	-2.55	96.45	94.94	-2.20	96.21	92.09	-2.50
GUARD (w=8)	94.77	90.92	-2.47	95.95	93.90	-2.15	95.82	91.11	-2.43
GUARD (w=9)	95.62	89.44	-2.55	96.59	92.31	-2.20	96.04	91.50	-2.48

Table 11: Llama-3: The performance under different datasets and w values shows that w = 7 is the appropriate locality window.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
GUARD (w=2)	94.87	92.52	-2.58	96.68	91.68	-2.20	96.25	90.89	-2.60
GUARD (w=3)	96.57	93.39	-2.71	97.49	91.89	-2.31	97.29	91.02	-2.70
GUARD (w=4)	95.10	91.64	-2.53	96.73	92.45	-2.19	96.10	91.26	-2.54
GUARD (w=5)	95.70	91.43	-2.66	97.32	89.66	-2.29	96.82	91.52	-2.65
GUARD (w=6)	94.97	91.14	-2.54	96.61	91.50	-2.21	96.20	91.57	-2.55
GUARD (w=7)	95.69	90.58	-2.63	96.85	90.49	-2.27	96.78	94.56	-2.64
GUARD (w=8)	95.05	89.26	-2.21	96.56	92.41	-2.21	96.20	92.64	-2.57
GUARD (w=9)	95.57	92.42	-2.63	97.07	93.88	-2.26	96.65	89.73	-2.62

Table 12: Mistral-v0.3: The performance under different datasets and w values shows that w = 7 is the appropriate locality window.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
GUARD (w=2)	95.32	89.45	-2.65	96.71	87.45	-2.16	96.28	89.39	-2.35
GUARD (w=3)	96.81	90.16	-2.62	97.53	91.44	-2.27	95.46	90.41	-2.43
GUARD (w=4)	95.86	88.22	-2.47	96.60	89.90	-2.14	96.88	89.43	-2.56
GUARD (w=5)	96.65	88.47	-2.56	97.33	89.22	-2.22	95.26	91.65	-2.67
GUARD (w=6)	95.46	90.56	-2.46	96.70	92.74	-2.15	96.44	90.16	-2.52
GUARD (w=7)	96.54	88.16	-2.55	97.23	90.76	-2.22	96.97	91.29	-2.65
GUARD (w=8)	95.59	87.81	-2.49	96.71	91.62	-2.15	96.30	91.09	-2.35
GUARD (w=9)	96.20	89.79	-2.54	97.09	88.50	-2.22	95.69	90.98	-2.64

Table 13: Gemma-7B: The performance under different datasets and w values shows that w = 7 is the appropriate locality window.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
GUARD (w=2)	96.64	84.21	-2.79	96.79	91.26	-2.48	97.29	84.71	-2.76
GUARD (w=3)	96.64	84.21	-2.79	96.79	91.26	-2.48	97.29	84.71	-2.76
GUARD (w=4)	94.39	83.56	-2.49	95.86	87.44	-2.25	95.56	83.13	-2.51
GUARD (w=5)	96.28	83.06	-2.66	96.49	89.83	-2.36	96.66	83.37	-2.64
GUARD (w=6)	94.60	79.72	-2.50	95.92	91.78	-2.26	95.69	81.43	-2.52
GUARD (w=7)	95.64	85.55	-2.62	96.19	90.61	-2.35	96.35	85.83	-2.62
GUARD (w=8)	94.69	84.25	-2.53	95.50	90.60	-2.55	95.68	82.50	-2.53
GUARD (w=9)	95.56	78.64	-2.62	96.17	89.35	-2.35	96.24	80.30	-2.61

Table 14: GPT2-xl: The performance under different datasets and w values shows that w = 7 is the appropriate locality window.

C Pseudo Code Design

Algorithm 1 GUARD Algorithm

Input:
Prompt, Model, MaxTokens, w (window size), λ (decay factor)
Output:
Generated text continuation

function GENERATETEXT(Prompt, Model, MaxTokens, w, λ)
  $Output \leftarrow$ Prompt
  $entropy\_history \leftarrow [\,]$; $tokencounts \leftarrow \{\}$; $\varepsilon \leftarrow 1e-6$
  for $t = 1$ to $MaxTokens$ **do**
    $distribution \leftarrow$ Model.getDistribution(Output)
    $V \leftarrow$ distribution.size
    $H_{loc} \leftarrow -\sum$(distribution $*$ log(distribution)) // Eq. 2
    $entropy\_history[t] \leftarrow H_{loc}$
    if $t < w$ **then**
      $\delta_{global} \leftarrow$ arctanh$((H_{loc} -$ med$(entropy_history))/\log|V|)$ //Eq. 9
      $\delta_{loc} \leftarrow 0$
    else
      $T \leftarrow$ len$(entropy_history)$
      $numerator \leftarrow \sum_{i=1}^{T}(\lambda^{T-i} * entropy\_history[i])$
      $denominator \leftarrow \sum_{i=1}^{T}(\lambda^{T-i})$
      $H_{glob,T} \leftarrow numerator/denominator$ //Eq. 1
      $med\_H\_recent \leftarrow$ median$(entropy_history[T-w:T])$
      $med\_diff \leftarrow med\_H\_recent -$ med$(H_{glob,T})$
      $r_{change} \leftarrow |H_{loc} - entropy\_history[T-1]|/(entropy\_history[T-1]+\varepsilon)$
      $r_{diff} \leftarrow |med_diff|/(med(H_{glob,T})+\varepsilon)$
      $q \leftarrow 1.0 + r_{change} + r_{diff}$ //Appendix B
      $\delta_{loc} \leftarrow q *$ arctanh$((H_{loc} - med_H_recent)/\log|V|)$ //Eq. 8
      $\delta_{global} \leftarrow q *$ arctanh$(med_diff/\log|V|)$ //Eq. 9
    end if
    $\lambda_k \leftarrow |\delta_{loc}|/(|\delta_{loc}| + |\delta_{global}| + \varepsilon)$ //Eq. 10
    $k_{signal} \leftarrow \exp(\lambda_k * \delta_{loc} + (1-\lambda_k) * \delta_{global})$
    $k_t \leftarrow 10 * (1 - 1/(k_{signal}+1)) + 5$ //Eq. 7
    $k_{signal\_\alpha} \leftarrow \exp(\lambda_k \delta_{loc} \log|k_t| + (1-\lambda_k)\delta_{global}\log|k_t|)$
    $\alpha_t \leftarrow 1 - 1/(k_{signal\_\alpha}+1)$ //Eq. 12
    $candidates \leftarrow$ Select top-k_t candidates
    $scores \leftarrow \{\}$
    for each token v in candidates **do**
      $count \leftarrow$ tokencounts.get$(v, 0)$
      $scores[v] \leftarrow$ distribution$[v] * (\alpha_t^{\text{count}})$ //Eq. 11
    end for
    $next\_token \leftarrow$ argmax$(scores)$
    $tokencounts[next\_token] \leftarrow$ tokencounts.get$(next_token, 0) + 1$
    $Output \leftarrow Output + next\_token$
  end for
  return $Output$[len(Prompt):]
end function

D Metrics

Diversity

This metric aggregates n-gram repetition rates:

$$\mathrm{DIV} = \prod_{n=2}^{4} \frac{|\text{ unique n-grams } (\mathrm{x_{cont}})\,|}{\text{total n-grams } (\mathrm{x_{cont}})\,|}.$$

A low diversity score suggests the model suffers from repetition, and a high diversity score means the model-generated text is lexically diverse.

MAUVE

MAUVE (Pillutla et al., 2021) is a metric designed to quantify how closely a model distribution Q matches a target distribution P of human texts. Two main types of error contribute to any discrepancy between Q and P:

- **Type I Error:** Q assigns high probability to text that is unlikely under P.
- **Type II Error:** Q fails to generate text that is plausible under P.

These errors can be formalized using the Kullback–Leibler (KL) divergences $\mathrm{KL}(Q \,\|\, P)$ and $\mathrm{KL}(P \,\|\, Q)$. If P and Q do not share the same support, at least one of these KL divergences will be infinite. To address this issue, Pillutla et al. (2021) propose measuring errors through a mixture distribution

$$R_\lambda \;=\; \lambda P \;+\; (1-\lambda)\, Q \quad \text{with } \lambda \in (0,1).$$

This leads to redefined Type I and Type II errors given by

$$\mathrm{KL}\big(Q \,\|\, R_\lambda\big) \quad \text{and} \quad \mathrm{KL}\big(P \,\|\, R_\lambda\big),$$

respectively.

By varying λ and computing these two errors, one obtains a *divergence curve*

$$\mathcal{C}(P,Q) \;=\; \big\{\big(\exp\!\big(-c\,\mathrm{KL}(Q \,\|\, R_\lambda)\big),\; \exp\!\big(-c\,\mathrm{KL}(P \,\|\, R_\lambda)\big)\big) \;:\; R_\lambda = \lambda P + (1-\lambda) Q,\; \lambda \in (0,1)\big\},$$

where $c > 0$ is a hyperparameter that controls the scaling.

Finally, $\mathrm{MAUVE}(P,Q)$ is defined as the area under the divergence curve $\mathcal{C}(P,Q)$. Its value lies between 0 and 100, with higher values indicating that Q is more similar to P.

Coherence

Proposed by Su et al. (2022), the coherence metric is defined as the averaged log-likelihood of the generated text conditioned on the prompt as:

$$\mathrm{Coherence}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{|\hat{\boldsymbol{x}}|} \sum_{i=1}^{|\hat{\boldsymbol{x}}|} \log p_{\mathcal{M}}\left(\hat{\boldsymbol{x}}_i \mid [\boldsymbol{x} : \hat{\boldsymbol{x}}_{<i}]\right)$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are the prompt and the generated text, respectively; [:] is the concatenation operation and $\mathcal{M}$ is the OPT model (2.7B) (Zhang et al., 2022).

BERTScore

BERTScore has shown positive correlations with human judgements (Zhang et al., 2019), therefore, we propose using it as an additional metric for evaluating the quality of the generated texts by each generation method with respect to the corresponding reference texts provided by human experts. By using contextual embeddings, BERTScore (Zhang et al., 2019) computes a similarity score for each token in the candidate sentence with each token in the reference sentence.

E Human Evaluation

To reflect the share of human raters favoring each decoding strategy, as depicted in Table 15 in the main paper, we apply the following scoring approach:

$$\text{Score}_{\text{ACS}} = \frac{\#(\text{ACS is better}) + 0.5 \times \#(\text{tie})}{\#(\text{ACS is better}) + \#(\text{tie}) + \#(\text{GUARD is better})}$$

$$\text{Score}_{\text{GUARD}} = 100 - \text{Score}_{\text{ACS}}.$$

Dataset	Coherence			Fluency		
	ACS is better	ACS and GUARD are similar	GUARD is better	ACS is better	ACS and GUARD are similar	GUARD is better
Wikitext	19%	37%	44 %	11%	67%	23%
Wikinews	24%	33%	44%	12%	67%	22%
BookCorpus	36%	27%	38%	18%	69%	13%
All	26%	32%	42%	14%	67%	19%

Table 15: Human evaluation results for ACS vs. GUARD across different datasets. Text generations are rated based on their semantic coherence and fluency.

Further, we measure the inter-rater agreement by computing Fleiss' kappa:

$$\kappa_{Fleiss} = 0.41$$

which reflects a moderate agreement across the human evaluators.

General instructions

-Below are multiple prompts and stories (narratives) generated from two different methods.
-Please note that the assignment to "Method A" or "Method B" is random, so each column includes examples of different methods
-Please evaluate both the coherence and the fluency of the methods

Important definitions of evaluation criteria

-Coherence: The story feels like one consistent story, and not a bunch of jumbled topics. Stays on-topic with a consistent plot, and doesn't feel like a series of disconnected sentences.
-Fluency: The story is written in grammatical English. No obvious grammar mistakes that a person wouldn't make.
An incomplete final word or incomplete sentence does not count as a mistake and should not affect fluency. The English sounds natural. Note: do not take off points for spaces between pu
(e.g. "don 't") and simpler sentences. Simple English is as good as complex English, as long as everything is grammatical.

Please read the prompt and the two possible continuations generated by Method A and B

Please select one of the three options by entering a value of 1 in your option of choice

Coherence evaluation: You have evaluated 0 out of 240 stories

Fluency evaluation: You have evaluated 0 out of 240 stories

Generation examples				Coherence			Fluency		
ID	Prompt	Method A	Method B	A is more coherent	A and B are similar	B is more coherent	A is more fluent	A and B are similar	B is more fluent
1				0	0	0	0	0	0
2				0	0	0	0	0	0
3				0	0	0	0	0	0
4				0	0	0	0	0	0
5				0	0	0	0	0	0
6				0	0	0	0	0	0
7				0	0	0	0	0	0
8				0	0	0	0	0	0
9				0	0	0	0	0	0
10				0	0	0	0	0	0
11				0	0	0	0	0	0
12				0	0	0	0	0	0
13				0	0	0	0	0	0
14				0	0	0	0	0	0
15				0	0	0	0	0	0

Figure 3: Human evaluation form, including general instructions and definitions for the evaluation criteria.

F Prompt design for LLM-as-a-judge

Instructions

You are a language evaluation specialist tasked with conducting pairwise comparisons of machine-generated texts. You will compare outputs from two different decoding methods (A and B) that were generated using the same input prompt.

Evaluation Process

For each text pair, assess which version demonstrates superior quality according to six specific metrics. Your evaluation should result in one of three judgments for each metric:

- **Method A is better**
- **Both methods are similar**
- **Method B is better**

Important Note The underlying decoding methods are randomized in their labeling as "Method A" or "Method B" across different evaluations. Do not attempt to identify patterns based on these labels, as they are arbitrarily assigned for each comparison.

Evaluation Metrics

1. Fluency Measures how natural, smooth, and grammatically correct the text reads. Evaluates whether the language flows naturally without awkward phrasing, grammatical errors, or unnatural constructions.

2. Coherence Assesses logical connection between ideas and overall text organization. Evaluates whether the text maintains consistent themes, follows logical progression, and avoids contradictions or non-sequiturs.

3. Factuality Measures accuracy and truthfulness of factual claims. Evaluates whether information presented is correct and free from errors, fabrications, or misrepresentations.

4. Informativeness Assesses the substantive content and value of the information provided. Evaluates whether the text delivers meaningful, relevant content rather than being vague, repetitive, or content-poor.

5. Interestingness Measures how engaging, compelling, or captivating the text is. Evaluates whether the content holds attention through creativity, unique insights, or engaging stylistic elements.

6. Story Development Assesses how effectively the narrative unfolds and progresses (where applicable). Evaluates character development, plot progression, pacing, and overall narrative structure in story-based texts.

G Further Comparative Experiments

Different Models We examine the impact of different models on the quality of generated text, such as Llama-2, Gemma-7B, Llama-3.1, Mistral-v0.3, and Deepseek-llm-7B-base across three datasets: Wikinews, Wikitext, and BookCorpus. The results are detailed in Table 16 (diversity, MAUVE, and coherence) and Table 17 (BERTScore) reveal notable differences between the models.

Dataset	Model	Diversity			MAUVE			Coherence		
		CS	ACS	GUARD	CS	ACS	GUARD	CS	ACS	GUARD
Wikitext	Gemma-7B	68.99	47.99	**96.54**	85.05	68.45	**88.16**	-1.50	-1.33	**-2.55**
	Llama-3.1	9.0	81.17	**95.84**	27.05	75.97	**91.25**	-0.74	-1.52	**-2.55**
	Mistral-v0.3	**97.26**	89.82	96.59	85.19	81.59	**90.58**	-2.09	-1.56	**-2.63**
	Deepseek-base-7B	77.43	32.14	**96.55**	56.26	46.92	**88.96**	**-2.49**	-1.23	-2.66
Wikinews	Gemma-7B	83.73	72.81	**97.23**	87.26	86.34	**90.76**	-1.52	-1.29	**-2.22**
	Llama-3.1	46.84	89.66	**96.45**	82.36	69.44	**94.94**	-1.25	-1.66	**-2.20**
	Mistral-v0.3	**98.25**	93.42	96.85	83.89	80.75	**90.49**	-1.92	-1.63	**-2.27**
	Deepseek-base-7B	94.13	61.96	**97.14**	61.34	73.39	**94.84**	-3.69	-1.66	**-2.23**
BookCorpus	Gemma-7B	65.65	39.76	**96.97**	86.65	70.59	**91.29**	-1.55	-1.13	**-2.65**
	Llama-3.1	16.45	84.86	**96.21**	52.94	78.09	**92.09**	-1.00	-1.64	**-2.50**
	Mistral-v0.3	**97.27**	90.74	96.78	79.03	75.50	**94.56**	-2.29	-1.69	**-2.64**
	Deepseek-base-7B	62.33	42.61	**96.86**	62.96	47.28	**90.07**	-2.41	-1.24	**-2.61**

Table 16: Comparing GUARD with CS ($k = 10, \alpha = 0.6$) and ACS ($q = 1$) across datasets and models of varying size. In Table 17, we also report the results of the influence of model architectures. Best results in **bold**.

In Table 17, we also investigate the influence of model architectures on the performance of representative contrastive-search-based decoding methods.

Dataset	Model	CS BERTScore (%)	ACS BERTScore (%)	GUARD BERTScore (%)
Wikitext	Llama-2	81.13	80.59	**81.50**
	Gemma-7B	**81.76**	81.61	81.73
	Llama-3.1	80.89	80.98	**81.76**
	Mistral-v0.3	81.14	81.25	**81.84**
	Deepseek-base-7B	80.00	80.02	**81.30**
Wikinews	Llama-2	82.00	82.23	**83.39**
	Gemma-7B	**83.54**	83.72	83.34
	Llama-3.1	**83.89**	82.25	83.76
	Mistral-v0.3	82.76	83.08	**83.52**
	Deepseek-base-7B	77.28	80.03	**83.32**
BookCorpus	Llama-2	79.74	81.36	**81.92**
	Gemma-7B	81.17	80.82	**81.61**
	Llama-3.1	81.12	81.07	**81.93**
	Mistral-v0.3	**81.56**	81.43	81.83
	Deepseek-base-7B	79.87	80.04	**81.72**

Table 17: Comparison of CS ($k = 10, \alpha = 0.6$), ACS ($q = 1$), and GUARD in terms of BERTScore across different architectures/LLMs of varying sizes. Best results in **bold**.

Hyperparameter Choice for Competing Decoding Strategies We validate our choice of hyperparameters for the competing methods (Beam search, Top-k/Top-p sampling) in Tables 18, 19, and 20.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
B=3	15.24	32.26	-0.81	25.78	58.02	-0.91	5.48	20.23	-0.59
B=5	14.42	26.39	-0.77	24.94	54.31	-0.87	5.39	18.69	-0.56
B=10	12.88	23.26	-0.72	19.28	40.24	-0.77	4.02	14.17	-0.49
B=15	12.23	20.30	-0.70	16.89	34.88	-0.72	3.32	11.93	-0.46
B=20	11.52	17.63	-0.67	15.45	34.71	-0.69	2.94	10.68	-0.44
B=50	8.28	13.93	-0.61	9.7	22.19	-0.58	2.24	7.60	-0.45
GUARD (Ours)	92.86	90.82	-2.61	95.20	93.60	-2.38	96.18	92.59	-2.52

Table 18: Averaged automatic evaluation results for Qwen2.5-7B for beam search (with GUARD) with different beam-widths.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
k=3	59.01	82.40	-1.33	74.21	91.46	-1.38	47.18	87.86	-1.44
k=5	70.69	89.23	-1.5	82.74	94.65	-1.52	67.99	90.42	-1.72
k=10	77.30	82.91	-1.68	88.18	93.40	-1.69	82.67	91.36	-2.01
k=15	79.66	88.95	-1.77	89.83	95.30	-1.78	85.99	94.87	-2.17
k=20	81.62	89.73	-1.84	90.17	94.11	-1.84	88.14	92.96	-2.56
k=50	82.69	88.97	-1.98	92.13	94.52	-2.01	91.54	93.24	-2.53
GUARD (Ours)	92.86	90.82	-2.61	95.20	93.60	-2.38	96.18	92.59	-2.52

Table 19: Averaged automatic evaluation results for Qwen2.5-7B for Top-k sampling (with GUARD) with different k.

Method	Wikitext			Wikinews			BookCorpus		
	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.	div. (%)	MAUVE (%)	coh.
p=0.60	52.43	79.56	-1.43	69.02	90.38	-1.54	38.21	80.55	-1.34
p=0.70	64.30	83.92	-1.60	77.56	93.74	-1.65	56.81	85.18	-1.58
p=0.80	73.80	83.80	-1.77	84.98	96.08	-1.83	72.88	90.73	-1.81
p=0.90	79.72	88.62	-1.97	89.96	96.67	-2.02	84.76	93.30	-2.12
p=0.95	82.47	86.37	-2.11	91.47	96.09	-2.15	88.92	94.81	-2.29
GUARD (Ours)	92.86	90.82	-2.61	95.20	93.60	-2.38	96.18	92.59	-2.52

Table 20: Averaged automatic evaluation results for Qwen2.5-7B for Top-p sampling (with GUARD) with different p.

H Case Study: Generation Examples

Prompt	*Returning to Australia, Headlam became Deputy Chief of the Air Staff (DCAS) on 26 January 1965. He was appointed a Companion of the Order of*
Reference	the Bath (CB) " in recognition of distinguished service in the Borneo Territories " on 22 June. His tenure as DCAS coincided with the most significant rearmament program the Air Force had undertaken since World War II, and with manpower shortages stemming from this expansion and from Australia's increasing involvement in the security of South East Asia.
CS	St Michael and St George (CMG) in the 1966 New Year Honours. What is the sentiment of this tweet? To determine the sentiment of this tweet, I will analyze the content and context of the information provided.
ACS	St Michael and St George (CMG) on 26 June 1965. Promoted to Air Commodore on 1 July 1965, Headlam relinquished his appointment as DCAS on 15 July 1965. He was appointed Air Officer Commanding RAAF Base Amberley on 15 July 1965, a position he held until his retirement on 15 July 1967.
GUARD	the Bath (CB) in the 1965 New Year Honours. In May 1967, he moved to London as Chief Executive of the Civil Air Lines Branch, Department of Transport. Headlam became Air Commodore-in-Chief on 30 June. On 8 November, he assumed his third and last staff command when appointed Commander-in-Chief, Far East Region. His last appointment in July–August 1970 was the appointment as Chairman and Chief Executive Officer at Qantas Airways.

Table 21: Text continuations based on prompts from the Wikitext dataset. Deviations from the context are highlighted in red. The results reveal that GUARD generates contextually relevant and fluent narratives, whereas the CS method deviates from the intended context and ACS exhibits repetitive patterns. When combined with the automatic evaluation results.

Prompt	*Two people have died as the result of a road traffic accident involving a truck and a car in Devon, England. The emergency services were called to the scene of*
Reference	a collision involving a Vauxhall Corsa and a DAF lorry at 0830 GMT on Saturday. Both occupants of the Vauxhall, a man and a woman, were pronounced dead at the scene of the collision, which was on the A361 road in the town of Knowstone, near South Molton.
CS	the accident, which took place on the A379 at 1.30 p.m. Police are appealing for witnesses to come forward and help with their investigation. Anyone who saw the accident or has any information is asked to contact Devon and Cornwall Police on 101, quoting log number 466 of 20/07/2019.
ACS	the accident at Bovey Tracey, Torquay, at 13:30 (14:30 CEST) on Saturday. Devon and Cornwall Police said the driver of the car, a 24-year-old man, was pronounced dead at the scene.
GUARD	the accident on 8 October, near Okehampton in north Devon. Both occupants of the car, which had collided with the back of the truck, have since died.

Table 22: Text continuations based on prompts from the Wikinews dataset.

Prompt	*My eyes moved involuntarily from the pages of the book that I was no longer reading, to the faces of the girls who were now looking over at me. My*
Reference	heart began to pound in my chest, and I could hear the blood rushing in my ears. I could feel the heat of the color burning on my cheeks. I was embarrassed, no I was furious.
CS	heart skipped a beat as I realized that they were looking at me with a mix of curiosity and confusion.
ACS	eyes caught sight of one of the girls, and I couldn't help but stare. "Excuse me," one of the girls said, "but what are you looking at?" I looked up at her, and then back down at the book.
GUARD	heart beat faster as their gazes lingered upon me and, without even realising it myself, I was smiling back, my cheeks flushed and my eyes shining with happiness.

Table 23: Text continuations based on prompts from the BookCorpus dataset.