Adaptive Contrastive Search: Uncertainty-Guided Decoding for Open-Ended Text Generation

Esteban Garces Arias[1,2], Julian Rodemann[1], Meimingwei Li[1], Christian Heumann[1], Matthias Aßenmacher[1,2]

[1]Department of Statistics, LMU Munich, [2]Munich Center for Machine Learning (MCML)

Correspondence: esteban.garcesarias@stat.uni-muenchen.de

Abstract

Despite the remarkable capabilities of large language models, generating high-quality text remains a challenging task. Numerous decoding strategies—such as beam search, sampling with temperature, top-k sampling, nucleus (top-p) sampling, typical decoding, contrastive decoding, and contrastive search—have been proposed to address these challenges by improving coherence, diversity, and resemblance to human-generated text. In this study, we introduce Adaptive Contrastive Search (ACS), a novel decoding strategy that extends contrastive search (CS) by incorporating an adaptive degeneration penalty informed by the model's estimated uncertainty at each generation step. ACS aims to enhance creativity and diversity while maintaining coherence to produce high-quality outputs. Extensive experiments across various model architectures, languages, and datasets demonstrate that our approach improves both creativity and coherence, underscoring its effectiveness in text-generation tasks. We release our code, datasets, and models to facilitate further research.

1 Introduction

The Transformer (Vaswani et al., 2017) plays a key role in various generative natural language processing (NLP) tasks, such as generating stories, completing contextual text, and dialogue systems. However, the conventional method of training these models using maximum likelihood estimation (MLE) and decoding to the most probable sequence often results in substantial shortcomings. This can lead to repetitive and uncreative outputs, also known as *degenerate* text.

To address this issue, previous approaches have aimed to adjust the decoding strategy by incorporating sampling from less probable vocabularies. While this helps reduce repetitiveness, it introduces the problem of semantic inconsistency (Welleck et al., 2020). The sampled text may stray from or even contradict the original context provided by a human-written prompt.

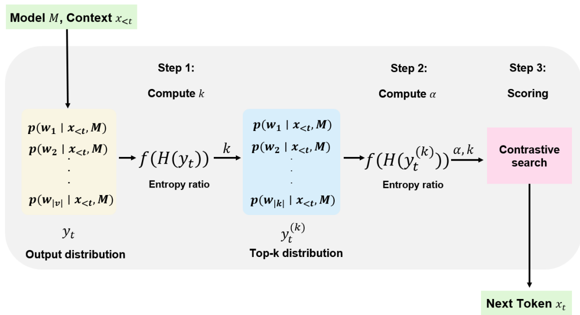


Figure 1: Visualization of the Adaptive Contrastive Search (ACS) process: A three-step procedure that uses entropy as a proxy for model uncertainty to automatically adjust contrastive search parameters.

In response, contrastive search (CS, Su et al., 2022) has been introduced. It employs a fixed balance between model confidence and degeneration penalty throughout the generation process, maintaining a blend of likelihood and diversity. However, it is important to note that this fixed weighting requires hyperparameter tuning and overlooks the unique demands of each generation step, where a different balance between model confidence and degeneration penalty might be desirable and even more effective.

We address this limitation by proposing adaptive contrastive search (ACS), an adaptive approach that automatically adjusts the hyperparameters of conventional CS. This method evaluates the model's uncertainty at each generation step and adjusts the weighting of both components without the need for manual intervention. The experimental outcomes demonstrate the effectiveness of our method, as it performs well in the task of open-ended text generation across different architectures, languages, and datasets.

Findings of the Association for Computational Linguistics: EACL 2024, pages 15060–15080

Contributions Our contributions can be summarized as follows:

1. We introduce an adaptive CS method based on the work by Su et al. (2022) that measures the uncertainty of the model at each time step to automatically adjust the number of candidate tokens and the degeneration penalty.

2. We conduct comprehensive experiments to compare our approach to various established decoding methods, such as nucleus sampling (Holtzman et al., 2019), contrastive decoding (CD, Li et al., 2023), and CS (Su et al., 2022), for open-ended text generation.

3. We offer new insights into MAUVE and its correlation with human judgments, highlighting the need for a more robust metric that better aligns with human preferences when evaluating decoding strategies for open-ended text generation.

4. Our code and datasets and results are publicly available under this link.

2 Related work

Decoding methods are generally categorized into two types: deterministic and stochastic.

Deterministic Methods. These approaches focus on choosing the text continuation with the highest probability according to the model's probability distribution. Prominent examples are beam search and greedy search. Recently, studies by Shao et al. (2017); Vijayakumar et al. (2018); Paulus et al. (2017); Klein et al. (2017) have demonstrated that solely maximizing the output probability frequently leads to degenerated or repetitive text sequences, a problem that has been addressed by stochastic, sampling-based methods.

Stochastic Methods. Top-k, proposed by Fan et al. (2018), samples from a subset of tokens $V^{(k)}$ that represent the tokens with the higher scores in the output distribution. Alternatively, nucleus-p sampling (Holtzman et al., 2019) samples from the smallest subset S with a total probability mass above a threshold p; specifically, S is the smallest subset such that the cumulative probability for tokens in S surpasses p. While these methods reduce model degeneration, the inherent stochasticity can lead to semantic divergence or disconnection from the human-written prompt.

To tackle the imbalance between coherence and diversity, methods such as typical sampling (Meister et al., 2023) and CD have been developed to produce more diverse and interesting text in open-ended settings. Typical sampling aims at generating based on the information content, which should be close to the expected information content, i.e., the conditional entropy of the model. CD, on the other hand, employs an expert language model (LM) and an amateur LM in parallel and searches for text that maximizes the difference between the expert's and the amateur's log probabilities, subject to plausibility constraints.

Our study, however, focuses on the work of Su et al. (2022), where they introduce Contrastive Search. In CS, given the prompt text $\boldsymbol{x}_{<t}$, the selection of the output token x_t follows:

$$x_t = \underset{v \in V^{(k)}}{\arg\max} \Big\{ (1-\alpha) \times \underbrace{p_\theta(v \mid \boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{\left(\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\}\right)}_{\text{degeneration penalty}} \Big\} \quad (1)$$

where $V^{(k)}$ is the set of top-k predictions from the LM's probability distribution $p_\theta(\cdot \mid \boldsymbol{x}_{<t})$. In Eq. (1), the first term, model confidence, is the probability of the candidate v predicted by the LM. The second term, degeneration penalty, measures how discriminative is the candidate v with respect to the previous context $\boldsymbol{x}_{<t}$ and $s(\cdot, \cdot)$ computes the cosine similarity between token representations. Intuitively, a larger degeneration penalty of v means it is more similar to the context, therefore more likely leading to undesirable repetitions in the generated output. The hyperparameter $k \in \mathbb{N}_{>0}$ determines the number of candidate tokens to be considered, while $\alpha \in [0, 1]$ regulates the importance of these two components.

Empirical results suggest different values for α and k, depending on the task, the datasets, and the language of interest, respectively (Su et al., 2022; Su and Xu, 2022; Su and Collier, 2023). An empirical study comparing CS and CD (Su and Collier, 2023) highlights the strengths and weaknesses of both approaches. The automatic evaluation results suggest that CD performs better on MAUVE Pillutla et al. (2021), while CS excels in diversity and coherence. Additionally, through extensive human evaluations, they demonstrate that human annotators universally prefer CS over CD by substantial margins. Given the contradictory results between MAUVE and human evaluations, their analysis reveals that balancing diversity and coherence metrics better correlates with human judgments.

Further studies have extended the concept of CS to incorporate additional criteria in scoring candidates for the next token. Chen et al. (2023) investigate the effect of adding a third criterion, fidelity, beyond model confidence and degeneration penalty to enhance the coherence of the generated text. This third criterion is again weighed by a hyperparameter β that is determined from empirical results. To the best of our knowledge, adaptive approaches based on the concept of CS have not been thoroughly explored.

3 Methodology

3.1 Incorporating Model Uncertainty

In this work, we propose an adaptive method that considers the estimated uncertainty of the model at time step t to automatically control k and α. In other words, our adaptive approach consists in modifying Eq. (1) as follows:

$$x_t = \underset{v \in V^{(k_t)}}{\arg\max} \Bigg\{ (1-\alpha_t) \times \underbrace{p_\theta(v \mid \boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha_t \times \underbrace{\left(\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\}\right)}_{\text{degeneration penalty}} \Bigg\} \quad (2)$$

where

$$k_t = 10 * \frac{\exp(\delta_t)}{\exp(\delta_t)+1} + 5 \quad (3)$$

with

$$\delta_t = q * \text{arctanh}\left(\frac{H(X)^{(t)} - \text{median}(H(X)^{(<t)})}{\text{maximum entropy}}\right) \quad (4)$$

and

$$\mathrm{H}(X)^{(t)} = -\sum_{x \in \mathcal{V}} p(x \mid \boldsymbol{x}_{<t}) \ln p(x \mid \boldsymbol{x}_{<t}). \quad (5)$$

Once k is selected, a similar procedure is followed to determine α_t:

$$\alpha_t = \frac{\exp(\delta_{t,k})}{\exp(\delta_{t,k})+1} \quad (6)$$

$$\delta_{t,k} = q * \text{arctanh}\left(\frac{H(X)^{(t,k)} - \text{median}(H(X)^{(<t,k)})}{\text{maximum entropy}^{(k)}}\right) \quad (7)$$

In other words, we follow a sequential procedure for k_t and α_t that involves these steps:

i) **Measuring uncertainty:** Compute the entropy of the output distribution denoted as $H(X)^{(t)}$.

ii) **Centering:** Subtract the median entropy of the previous prediction steps.

iii) **Scaling:** Divide by the maximum entropy. This step aims to obtain a relative measure, ensuring comparability across different vocabulary sizes.

iv) **Computation**: Pass the centered and rescaled entropy term through a sigmoid function, yielding the value of $\alpha_t \in (0,1)$ - or for the case of k - through a rescaled sigmoid function that yields positive integer values.

The scaling term *maximum entropy* refers to the entropy of a uniform distribution over a finite set $x_1, ..., x_{|\mathcal{V}|}$, where each token has an equal probability of $\frac{1}{|\mathcal{V}|}$. Consequently, this entropy remains constant over time. For a vocabulary of size $|\mathcal{V}|$, the maximum entropy is given by $\ln(|\mathcal{V}|)$, analogously, the maximum entropy for the distribution of the top-k tokens is given by $\ln(k)$.

Additionally, the parameter q serves as a temperature factor, influencing the range of k and α values at each time step. Adjusting q can either broaden or narrow this range: a lower temperature reduces variability, while a higher value allows for larger changes. This impact is demonstrated in Appendix A, Figures 3, 4, 5, 6, 7 and 8. However, it is important to note that our evaluation is based on a setup with no temperature (i.e., $q = 1$).

3.2 Theoretical Motivation

The general question that might arise is to why k_t and α_t should be chosen adaptively, i.e., why there is no global α nor a global k that is optimal at all time steps. Taking on a more statistical perspective on the problem offers an explanation: The degeneration penalty in constrastive search can be understood as a regularization term, see also (Chen et al., 2023). More precisely, we have:

Proposition 1 *The degeneration penalty* $\max_j\{s(h_v, h_{x_j})\}$ *is a function of the penalty* $||h_v - h_{x_j}||_2$ *in statistical Tikhonov-regularization, if the representations* h *are normalized.*

The proof can be found in Appendix B. Classical statistical regularization aims at preventing overfitting by smoothing out the effect of the training data on the model fit. In CS, the degeneration penalty plays a similar role: It attenuates the effect of the model on the chosen output token. Recall that we do not want to select tokens solely based on the model to prevent repetitive text generation. It is a well-known fact that optimal regularization parameters (corresponding to k_t and α_t here) for many statistical models correspond to the signal-to-noise ratio in the data, see e.g. the work by Kimeldorf and Wahba (1970); Rao et al. (2008); Hastie et al. (2009); Fahrmeir et al. (2022). The intuition is straightforward: For a given signal, the more noise in the data, the higher the optimal regularization parameter should be chosen to prevent overfitting to the latter. This motivates our adaptive approach to CS. Instead of choosing fixed k and α in the beginning, we choose it based on the observed variation of the object on which we want to prevent overfitting. In contrast to statistical modeling, however, this object is the model output, not the training data. We thus choose the optimal k_t and α_t based on the variation of the model output, i.e., its estimated uncertainty measured by δ_t and $\delta_{(t,k)}$, the standardized Shannon entropy. While several approaches to quantifying uncertainty exist, see Abdar et al. (2021) for an overview, we rely on the classical Shannon entropy since it is computationally efficient and tailored to measuring epistemic (reducible) *predictive* uncertainty, see (Hüllermeier and Waegeman, 2021, section 3.3), as required here.

4 Experimental Setup

In this section, we describe the metrics, datasets, baseline models, and human evaluation settings.

4.1 Evaluation Metrics

We follow Su and Xu (2022) and use three metrics to automatically measure the quality the generations: Diversity, MAUVE, and Coherence.

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) score measures the distribution similarity between the set of generated text and the set of gold references.

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

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

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).

Human Eval. In order to evaluate the quality of the generated text, we consider two critical aspects: fluency and coherence. A fluent piece of text is written in grammatical English and has a natural flow (e.g. excluding unnatural repetition or web formatting). A coherent piece of text should stay on topic with the prompt and avoid unnatural topic drift. We provide five native English speakers with 240 competing continuations (A and B) of the same prompt and ask them to rate their coherence and fluency. Definitions and instructions for the rating process are shown in Appendix C, Figure 9.

4.2 Datasets

Following previous studies, we evaluate our proposed method on three domains for open-ended text generation: news, Wikipedia articles, and stories. For the news domain, we use articles from Wikinews (2000 examples); for the Wikipedia domain, we use the WikiText-103 dataset (1314 examples; Merity et al., 2016); and for the story domain, we use the BookCorpus (Project Gutenberg split, 1947 examples; Zhu et al., 2015). Each of the examples contains a prompt and a gold reference i.e. human-generated continuation for evaluation purposes. We extract the prompts and decode 256 tokens for the continuations. Finally, we evaluate the generated text based on both the set of metrics (as described in Sec. 4.1) and human preferences.

4.3 Baselines

We compare ACS to widely used decoding methods (including deterministic and stochastic approaches): greedy search, beam search, top-k sampling, nucleus sampling, typical decoding, CD, and CS with constant $\alpha = 0.6$ and two versions of k: 5 and 10. We include the latter for a fair comparison: The CS generations with $k = 10$ achieve better automatic evaluation scores than those generated with

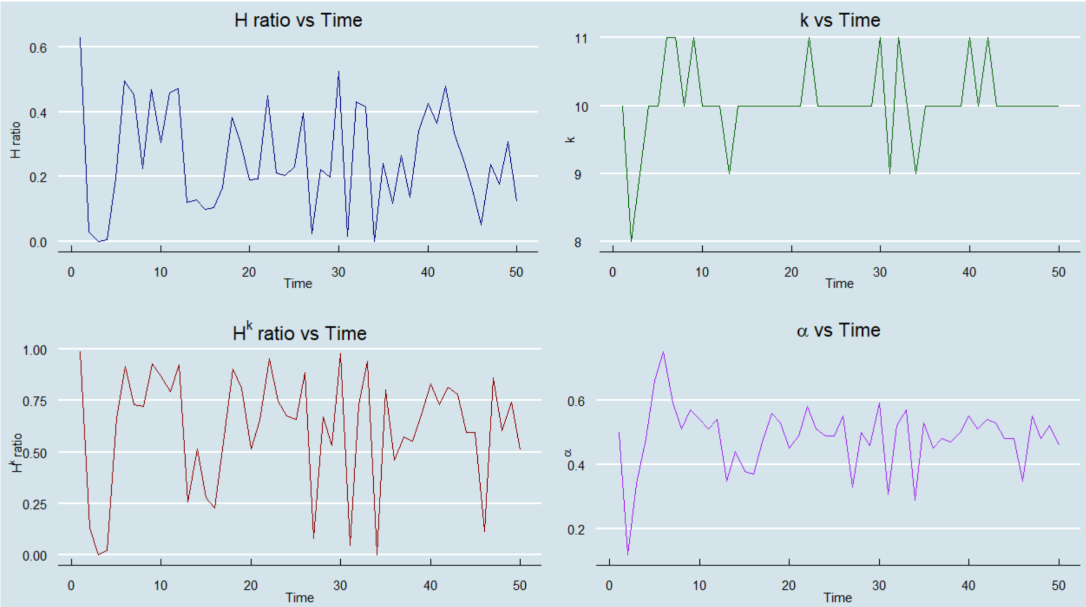


Figure 2: Visualization of uncertainty over time, measured by the Shannon entropy of the output distribution (first row, left). It is used to determine the value of k over time (right). The second row illustrates the entropy of the top-k tokens distribution, which is used to compute the value of α (right).

Prompt: *Knowing that she would be staying in, she started by choosing a pair of fitted, soft, black slippers, the type that barely covered her feet but gave*
Generated story: *her a sense of comfort. As she walked to the dining room, she took a moment to admire the decor and thought about what she would do for dinner. Her family was going to be here for a few days, and she wanted to make the most of the time they had before they left.*

Method	Wikinews			Wikitext			Story		
	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑
Greedy Search*	3.55	13.96	**-0.47**	1.77	4.91	**-0.41**	0.86	2.65	**-0.34**
Top- k Sampling*	91.56	89.86	-2.22	87.49	81.00	-2.37	91.22	87.49	-2.45
Nucleus Sampling*	93.54	89.45	-2.61	92.16	86.54	-3.03	94.50	91.47	-3.02
Typical Sampling*	95.37	90.97	-3.26	94.82	86.07	-3.71	96.29	88.58	-3.68
CD*	91.57	**92.20**	-2.16	88.02	**91.46**	-2.19	86.41	**93.17**	-2.09
CS ($k = 5, \alpha = 0.6$)	93.72	84.14	-1.39	89.35	77.97	-1.56	93.06	84.74	-1.61
CS ($k = 10, \alpha = 0.6$)	96.30	87.53	-1.73	94.09	77.97	-1.93	95.46	84.96	-1.91
ACS (Ours, $q = 1$)	95.22	79.45	-1.60	92.72	78.67	-1.74	93.89	80.72	-1.71
Bonus: *DoubleExp*	**97.39**	90.65	-2.12	**96.58**	84.07	-2.18	**97.37**	85.66	-2.16

Table 1: Automatic evaluation results: Numbers marked with * are obtained using the generated texts originally released by Li et al. (2023), CS results are taken from Su and Xu (2022). The highest scores are highlighted in **bold**.

$k = 5$. Furthermore, they are more comparable to our method, centered around $k = 10$. Further, we include an additional adaptive method: *DoubleExp*, which consists on an exponentiation of the argument of the sigmoid function, with the purpose of reaching values of α closer to 0 or 1. Our goal with this method is to exemplify discrepancies between human judgment and MAUVE. The respective human evaluation results are visualized in Table 2 and its implementation is described in Appendix F.

4.4 Models

We explore the relationship between model size and the effect of ACS. For this purpose, we use three open-source autoregressive models: gpt2-xl, gpt2-large, and gpt2-medium (Radford et al., 2019).

5 Results

5.1 Automatic evaluation results

The automatic evaluation of generated stories, based on diversity, MAUVE, and coherence are presented in Table 1. We observe that CS-based approaches tend to foster diversity, while having a slighter loss of coherence, compared to other decoding methods, such as CD and Typical sampling. An additional common trait is that CS-based approaches exhibit lower MAUVE scores, where CD

Dataset	Coherence			Fluency		
	CS is better	CS and DoubleExp are similar	DoubleExp is better	CS is better	CS and DoubleExp are similar	DoubleExp is better
Wikinews	56%	34%	10%	32%	58%	10%
Wikitext	34%	46%	20%	29%	63%	8%
Story	49%	31%	20%	32%	58%	10%
All	48%	36%	16%	28%	62%	10%

Dataset	Coherence			Fluency		
	CS is better	CS and ACS are similar	ACS is better	CS is better	CS and ACS are similar	ACS is better
Wikinews	24%	45%	31%	16%	67%	18%
Wikitext	37%	28%	35%	9%	74%	18%
Story	41%	27%	32%	21%	59%	21%
All	31%	32%	37%	17%	62%	21%

Table 2: Human evaluation results for CS vs. *DoubleExp* (top) and CS vs. ACS (bottom) across different datasets.

Prompt	*Butt criticized Donald's controls in certain situations in the game, as well as the difficulty of some levels and puzzles. Buchanan also criticized the controls, calling*	div.	coh.
Beam	them " too easy " and " too difficult ". He also criticized the game's graphics, stating that the game's graphics were " too simplistic " and "too simplistic ". He also criticized the game's graphics, stating that the game's graphics were " too simplistic" and " too simplistic ". He also criticized the graphics, stating that the game's graphics were " too simplistic " and " too simplistic ". He also criticized the graphics, stating that the game's graphics were " too simplistic " and " too simplistic ". He also criticized the graphics, stating that the game's graphics were"...	0.02	0.605
Nucleus	them " offensively rudimentary ". After the game was released, Buchanan wrote a number of game reviews about his review of the game, primarily summarizing what he thought the reviewers had enjoyed about it. Buchanan was pleased with how consistently these reviews awarded the game. For one, he praised the puzzles, compared them to Nos. 1 and 2 and enjoyed the construction of the artificial intelligence in the game's forms. He found no fault with the puzzles, even in those situations where they seemed repetitive. According to an Xbox Magazine review, " every time...	0.98	0.597
CD	them " unimpressive " and a " nightmare " of an experience to play with players unfamiliar with Tetris. On the other hand, his opinion was shared by other reviewers, and some were critical of the game's technical design for the Wii version of Tetris. In addition, Tintin's review included a quote from Roger Ebert, who said that Tetris was better than the original game due to its simplicity and ease of play. Ebert's comments were included in the game's DVD commentary, released on March 22,2010 . It is unclear if any of the video commentary was taken from ...	0.98	0.626
CS	them " unimpressive " and a " nightmare " of an experience to play with players unfamiliar with Tetris. On the other hand, his opinion was shared by other reviewers, and some were critical of the game's technical design for the Wii version of Tetris. In addition, Tintin's review included a quote from Roger Ebert, who said that Tetris was better than the original game due to its simplicity and ease of play. Ebert's comments were included in the game's DVD commentary, released on March 22,2010 . It is unclear if any of the video commentary was taken from ...	0.98	0.626
ACS (Ours, $q = 1$)	them " a pain in the ass to get used to." On the other hand, his opinion was shared by other reviewers, and some were critical of the game's technical design for the Wii version of Tetris. In addition, Tintin's review included a quote from Roger Ebert, who said that Tetris was better than the original game due to its simplicity and ease of play. Ebert's comments were included in the game's DVD commentary, released on March 22,2010 . It is unclear if any of the video commentary was taken from ...	0.98	0.629

Table 3: Case Study: Beam search produces degenerative repetitions while nucleus sampling produces text with incoherent semantics w.r.t. the prefix. Contrastive methods exhibit coherent and fluent text.

excels across all three datasets. We observe that the method *DoubleExp*, which consists on an exponentiation of the sigmoid arguments, provides a good balance of high diversity and MAUVE, maintaining coherence values that are lower than non-contrastive methods.

5.2 Human evaluation

The human evaluation scores for the CS and ACS methods, as well as for CS and *DoubleExp*, are displayed in Table 2. It is worth mentioning that high MAUVE scores do not always align with human judgments. For instance, evaluators consistently show a preference for CS over *DoubleExp* across all datasets. Conversely, ACS is favored in terms of fluency, and AC is preferred for coherence. Nonetheless, when considering all datasets together, there is a slight overall preference for ACS compared to its non-adaptive version.

5.3 Qualitative examples

We present qualitative examples to illustrate the distinct characteristics of different decoding strategies. Table 3 highlights the generated text variations, while Figure 2 visualizes the behavior of key parameters such as entropy, k, and α.

Method	Wikinews			Wikitext			Story		
	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑
ACS, q = 1	95.22	79.45	-1.6	92.72	78.67	-1.74	93.89	80.72	-1.71
ACS, q = 2	95.03	81.66	-1.57	92.69	77.48	-1.71	93.38	80.85	-1.67
ACS, q = 4	95.75	83.41	-1.76	94.02	81.56	-1.87	94.97	80.14	-1.82
ACS, q = 8	96.92	83.10	-2.02	95.23	77.79	-2.08	96.02	82.71	-2.04
ACS, q = 15	97.46	83.03	-2.24	96.39	81.66	-2.25	96.66	81.44	-2.23
ACS, q = 20	97.78	85.01	-2.32	96.55	81.61	-2.33	96.66	80.37	-2.26

Table 4: Ablation results for diversity, MAUVE, and coherence w.r.t. to the adaptiveness enforced by temperature q.

Method	**sec / story**↓	**# Tokens / sec**↑
CS (α = 0.6, k = 10)	11.6	21.98
ACS (q = 1)	15.7	16.29
ACS (q = 2)	15.9	16.14
ACS (q = 8)	16.3	15.35

Table 5: Comparison of generation speed for CS and ACS with different temperatures $q \in \{1, 2, 8\}$. Experiments were conducted with a GPU NVIDIA RTX 3090.

5.4 Ablation studies

To assess the impact of varying levels of adaptiveness, controlled by the temperature parameter, we conduct experiments for $q \in \{1 \text{ (no temperature)}, 2, 4, 8, 15, 20\}$. The results, summarized in Table 4, demonstrate the sensitivity of ACS to different values of q. As expected, increasing the temperature leads to higher diversity but at the cost of reduced coherence. Moreover, we observe that in both Wikitext and Story, the MAUVE score begins to decline as the generated texts become excessively diverse and erratic.

5.5 Generation speed

We have measured the average generation speed across all three datasets by varying the temperature $q \in \{1, 2, 8\}$ with respect to our baseline (CS with $k = 10$ and $\alpha = 0.6$). A summary is presented in Table 5, showing a decrease of 35% in the speed for ACS compared to CS with fixed α. A supplementary analysis of this for smaller values of k is provided in Appendix G.

5.6 Application to other languages

We evaluate the performance of our approach across eight additional languages: Arabic, Bengali, German, French, Hindi, Japanese, Dutch, and Chinese. For this, we utilize pre-trained GPT-2-based architectures of various sizes, comparing the Adaptive Contrastive Search (ACS) to standard Contrastive Search (CS). The detailed results are provided in Table 6. The scores vary greatly across languages and metrics, with particularly strong results in Bengali, French, Hindi, and Chinese. Nevertheless, a clear trend emerges: ACS, with the exception of German, consistently achieves comparable or superior MAUVE scores while maintaining a balance between coherence and diversity, outperforming its static counterpart.

5.7 Effect of varying model sizes

We examine the influence of model size on the quality of text generation, focusing on three different sizes: gpt2-xl (1.50B), gpt2-large (0.76B), and gpt2-medium (0.35B). The generated outputs are evaluated across the Wikinews, Wikitext, and Story datasets. As shown in Table 7, there are marked differences in performance, particularly with gpt2-medium, where the diversity under CS is substantially diminished across all datasets. We hypothesize that this is linked to the model's isotropy, where only high values of α can foster diversity in the generations. A visual inspection further supports this observation, as the outputs from gpt2-medium tend to be repetitive and degenerate, resembling those produced by greedy or beam search.

5.8 Findings about MAUVE

Our human evaluation results, as detailed in Table 2, corroborate the findings of Su and Xu (2022). The discrepancies observed indicate that MAUVE does not consistently align with human preferences. Specifically, the *DoubleExp* method yields higher MAUVE values compared to CS and ACS. However, human evaluators consistently rate *DoubleExp* as less coherent and fluent than the CS approach with $k = 10$ and $\alpha = 0.6$. Moreover, we identified two considerable issues related to varying the truncation length in pairwise sentence comparisons: (1) Inconsistent results that lead to different conclusions regarding the optimal decoding method, and (2) substantial differences in sample sizes, as illustrated in Appendix E, Table 8.

Language	Contrastive Search			Adaptive Contrastive Search			Δ		
	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.	div.(%)	MAUVE(%)	coh.
Arabic	89.55	70.53	-1.51	60.71	89.94	-1.23	-28.84	19.41	0.28
Bengali	72.48	89.87	-1.24	85.17	96.31	-1.34	12.69	6.44	-0.10
German	97.95	72.80	-2.16	93.04	42.82	-1.07	-4.91	-29.98	1.09
French	95.74	93.21	-2.27	92.49	96.41	-2.08	-3.25	3.20	0.19
Hindi	98.99	95.95	-1.00	98.90	92.99	-1.00	-0.09	-2.96	0.00
Japanese	50.47	72.69	-0.92	39.47	83.30	-1.80	-11.00	10.61	-0.88
Dutch	95.47	33.57	-2.96	98.03	72.32	-1.30	2.56	38.75	1.66
Chinese	91.42	93.28	-2.39	82.55	92.76	-2.26	-8.87	-0.52	0.13

Table 6: Comparison across different languages. Positive Δ-values indicate better performance of ACS vs CS.

Dataset	Model	Contrastive Search			Adaptive Contrastive Search			Δ		
		div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)	MAUVE(%)	coh.
Wikinews	gpt2-xl	93.72	88.14	-1.39	96.92	83.10	-2.02	3.20	-5.04	-0.63
	gpt2-large	93.80	78.55	-1.44	96.55	78.84	-2.06	2.75	0.29	-0.62
	gpt2-medium	3.66	12.86	-0.56	49.88	20.25	-6.22	46.22	7.39	-5.66
Wikitext	gpt2-xl	89.35	77.97	-1.56	95.23	77.79	-2.08	5.88	-0.18	-0.52
	gpt2-large	89.04	73.91	-1.59	95.67	80.00	-2.11	6.63	6.09	-0.52
	gpt2-medium	2.25	4.75	-0.47	64.13	10.91	-5.94	61.88	6.16	-5.47
Story	gpt2-xl	93.06	84.74	-1.61	96.02	82.71	-2.04	2.96	-2.03	-0.43
	gpt2-large	90.63	81.16	-1.56	95.82	80.42	-2.05	5.19	-0.74	-0.49
	gpt2-medium	1.22	3.08	-0.40	11.86	17.19	-6.13	10.64	14.11	-5.73

Table 7: Comparison of CS ($k = 10, \alpha = 0.6$) and ACS across different datasets and models of varying size.

5.9 Interpretability

We conducted additional experiments employing CS with varying values of k and α, measuring the same automatic metrics for human-generated text and examining which combinations of hyperparameters align most closely with the gold references. In Appendix D, our analysis reveals that extreme parameter settings, such as very low values of α, yield very low diversity and MAUVE scores but excessively high coherence (even in combination with high values of k). Conversely, very high values of α lead to texts that are overly diverse and incoherent (even with moderate values of k). A desired balance emerges from these observations: moderate values of k and α, such as $k = 10$ and $\alpha = 0.6$, tend to approximate human references by favoring both diversity and coherence. Our experiments show that ACS frequently generates results within this range, at times favoring either higher diversity or greater coherence, as dictated by the uncertainty-guided regularization.

6 Discussion and Future Work

In this study, we introduce a novel approach grounded in a CS framework. It is important to note, however, that the quality of generated text extends beyond just model confidence and diversity. Other factors, like informativeness, trustworthiness, and cohesion (as discussed by De Beaugrande and Dressler (1981)), contribute to the overall quality of the text. Moving forward, our aim is to expand our adaptable method to encompass these traits, thereby improving the evaluation of text quality through a more holistic approach. Furthermore, we would like to explore other criteria for the automatic selection of k and α. A path worth exploring could be related to the ratio between generation perplexity and model perplexity, where the algorithm would penalize deviations from the model perplexity through automatic adaptation. Finally, our study has centered on a specific task: Open-ended text generation. However, the potential influence of this adaptive decoding strategy on various tasks and contexts warrants further investigation. We aim to broaden our research to evaluate its efficacy in Machine Translation and Summarization, particularly within the scope of low-resource languages, where our approach may prove beneficial in scenarios where training examples are scarce. Finally, we wish to explore the potential of our approach beyond base models and analyze the effect after SFT and strategic prompting.

7 Conclusion

We introduce an uncertainty-guided adaptive method aimed at enhancing the quality of open-ended text generation outputs. Our approach calculates the estimated uncertainty at each time step using Shannon entropy and leverages this measure to dynamically adjust the weighting between model confidence and the degeneration penalty, as proposed by Su et al. (2022) and Su and Xu (2022). Our experiments demonstrate that this method performs well in terms of coherence and diversity, achieving MAUVE scores comparable to existing methods. It receives high ratings from human evaluators, particularly for the fluency of its outputs. This method requires no hyperparameter tuning and utilizes computational resources similar to the non-adaptive CS. Notably, unlike CD, it does not rely on two separate models, making it more versatile and suitable for various tasks. While it introduces some latency, our comprehensive studies comparing this approach to CS highlight its robustness across multiple languages and model sizes. We encourage further research into adaptive methods for open-ended text generation and advocate for the development of new metrics that better align with human judgments to improve the evaluation of decoding strategies.

Limitations

The proposed approach represents a potential step forward in enhancing text generation quality from language models. However, a key limitation lies in the method's narrow focus on two main objectives: model confidence and degeneration penalty. While these are critical for evaluating certain aspects of text quality, they do not fully capture the broad spectrum of desirable traits in generated text, such as informativeness, fluency, accuracy, trustworthiness, coherence, and cohesion. By not incorporating these additional factors into the decoding process, ACS may produce text that, while coherent and diverse, lacks depth or fails to effectively communicate complex or specialized information (and facts). Expanding the focus to include these dimensions could significantly improve the robustness and applicability of the method across diverse text generation tasks. Another limitation worth considering is the architectural choice of our experiments, which focuses on the family of gpt2-models, in particular in its gpt2-xl version. We plan to extend this analysis to more modern architectures, such as Mistral 7B, (Jiang et al., 2023, 2024), Llama2 7B (Touvron et al., 2023), Llama 3.1 8B (Dubey et al., 2024), Deepseek 7B (DeepSeek-AI et al., 2024), Qwen2 (Yang et al., 2024), Falcon2 (Malartic et al., 2024).

Additionally, the application of ACS has primarily been explored in the context of text generation tasks such as language modeling and natural language understanding. Its effectiveness on other tasks, such as machine translation, summarization, or multi-modal tasks, remains largely untested. Each of these tasks poses unique challenges and requirements in terms of text diversity, context preservation, and semantic accuracy. Investigating the adaptability and performance of the approach across a broader range of tasks will be essential for its broader generalizability. Lastly, while various measures of local uncertainty have been explored in the context of ACS, such as KL divergence, variance, perplexity, and entropy, there is still much room for exploration and refinement. Further experimentation and analysis are needed to determine the optimal combination of uncertainty metrics that can reliably produce high-quality, open-ended text generation across diverse domains and languages. However, the modularity of ACS in terms of choosing the δ-functions allows for a seamless further exploration of different approaches and applications.

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 commit to sharing our code, data, and other relevant resources to enhance reproducibility and further advancements in the field.

Acknowledgments

We wish to express our gratitude to Daniele Pugno and Nicolò Campagnoli for their technical support and visualizations. Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of BERD@NFDI, under grant number 460037581. Julian Rodemann acknowledges support by the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics" as well as by the Bavarian Institute for Digital Transformation (bidt) and the Bavarian Academy of Sciences (BAS) within a graduate scholarship.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen, and Chung-Chi Chen. 2023. Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation. *arXiv preprint arXiv:2310.14981*.

Robert-Alain De Beaugrande and Wolfgang U Dressler. 1981. *Introduction to text linguistics*, volume 1. longman London.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, 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, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, 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, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, 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, 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, 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, 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, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, 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, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, 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, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, 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, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, 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, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, 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, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, 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, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, 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, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, 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, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, 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, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D Marx. 2022. Regression models. In *Regression: Models, methods and applications*, pages 23–84. Springer.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *Preprint*, arXiv:1805.04833.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and

Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

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, 2024. Mistral 7b. *Preprint*, arXiv:2310.06825.

George S Kimeldorf and Grace Wahba. 1970. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *Preprint*, arXiv:1701.02810.

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. *Preprint*, arXiv:2210.15097.

Quentin Malartic, Nilabhra Roy Chowdhury, Ruxandra Cojocaru, Mugariya Farooq, Giulia Campesan, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Maksim Velikanov, Basma El Amel Boussaha, Mohammed Al-Yafeai, Hamza Alobeidli, Leen Al Qadi, Mohamed El Amine Seddik, Kirill Fedyanin, Reda Alami, and Hakim Hacid. 2024. Falcon2-11b technical report. *Preprint*, arXiv:2407.14885.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Preprint*, arXiv:2202.00666.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *Preprint*, arXiv:1705.04304.

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, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

C Radhakrishna Rao, H Toutenburg, Shalabh, and C Heumann. 2008. *Linear models and generalizations: Least squares and alternatives*, volume 3.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. *Preprint*, arXiv:1701.03185.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation. *Preprint*, arXiv:2210.14140.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Preprint*, arXiv:2202.06417.

Yixuan Su and Jialu Xu. 2022. An empirical study on contrastive search and contrastive decoding for open-ended text generation. *Preprint*, arXiv:2211.10797.

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, 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. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. *Preprint*, arXiv:1610.02424.

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho.

2020. Consistency of a recurrent language model with respect to incomplete decoding. *Preprint*, arXiv:2002.02492.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Preprint*, arXiv:1506.06724.

Appendix

A Example Generations

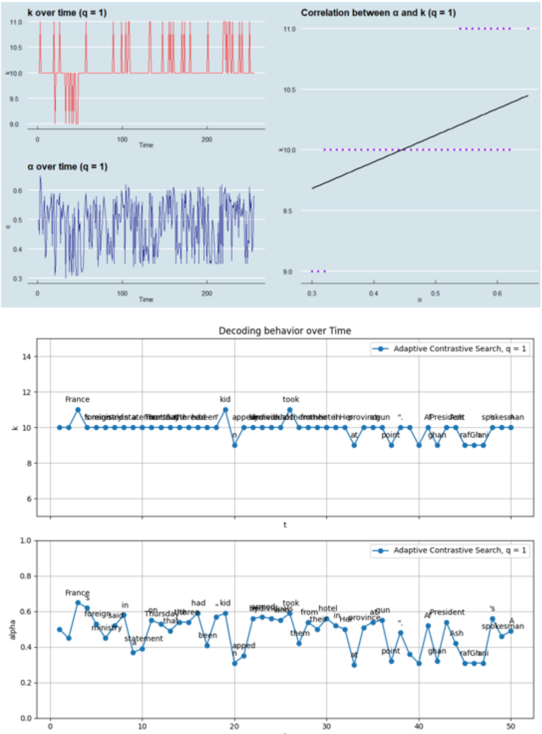


Figure 3: Decoding behavior over time from a Wikinews prompt, $q = 1$.

Prompt: Two French journalists kidnapped on Wednesday north-east of the capital Kabul are believed to be in good health, although the condition of their three Afghan assistants is unknown.
Generated story: France's foreign ministry said in a statement on Thursday that the three had been "kidnapped by armed individuals who took them from their hotel in Herat province at gunpoint". Afghan President Ashraf Ghani's spokesman Aimal Faizi said a helicopter had been sent to pick up the journalists and they were on their way to Bagram airbase in the north-centre of the country. Ghani, who took office in September, has faced criticism for failing to make progress in a war that has claimed more than 2,000 lives since US-led forces overthrew the Taliban in 2001. The United Nations Assistance Mission in Afghanistan (Unama) said this week that security in the country had deteriorated, with attacks on aid workers and foreign military convoys on the rise. Last month, Taliban fighters ambushed a convoy of UN peacekeepers in Kunduz province, killing at least 15 people and wounding more than 30. (FRANCE 24 with AFP)

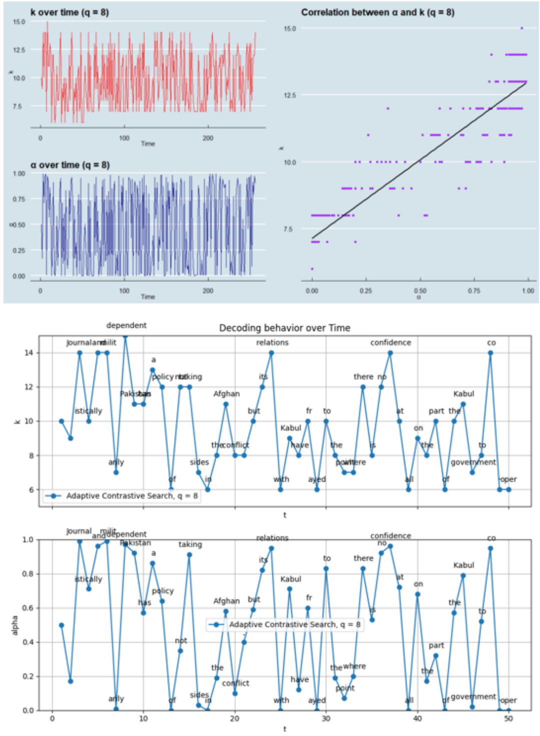


Figure 4: Decoding behavior over time from a Wikinews prompt, $q = 8$.

Prompt: Two French journalists kidnapped on Wednesday north-east of the capital Kabul are believed to be in good health, although the condition of their three Afghan assistants is unknown.
Generated story: Journalistically and militarily dependent Pakistan has a policy of not taking sides in the Afghan conflict, but its relations with Kabul have frayed to the point where there is no confidence at all on the part of the Kabul government to co-operate with Islamabad. Afghan President Hamid Karzai last week cancelled a high-level visit to the United States in protest at Washington's refusal to hand over Bin Laden on humanitarian grounds. Bin Laden escaped from Tora Bora four years ago by flying across the Hindu Kush mountains of Central Asia with help from al-Qaeda-linked guerrillas known as the Haqqani network. Since then his hideout has been under watch by U.S. Special Forces who have set up a command centre close to the Afghanistan-Pakistan border for intelligence, surveillance and counter-terrorism operations.

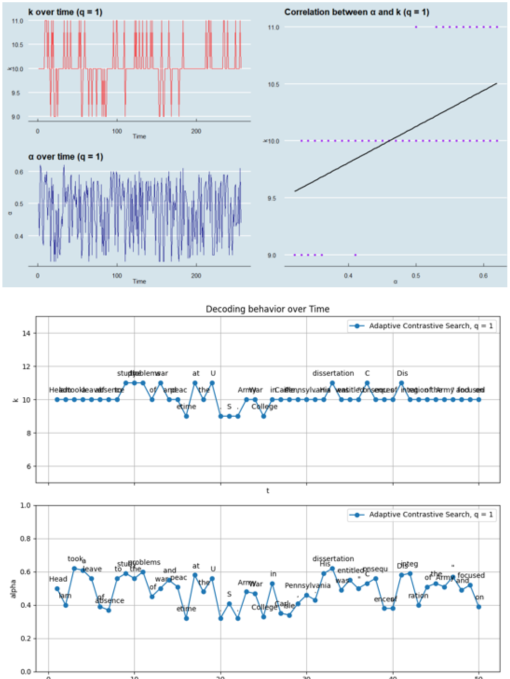


Figure 5: Decoding behavior over time from a Wikitext prompt, $q = 1$.

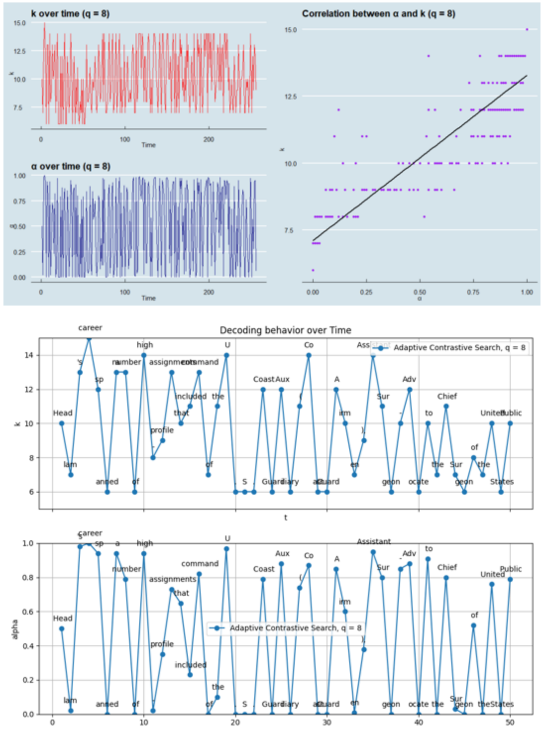


Figure 6: Decoding behavior over time from a Wikitext prompt, $q = 8$.

Prompt: Headlam served as Officer Commanding North @-@ Western Area in 1946, and as Director of Training from 1947 to 1950. In 1950 – 51,
Generated story: Headlam took a leave of absence to study the problems of war and peacetime at the U.S. Army War College in Carlisle, Pennsylvania. His dissertation was entitled "Consequences of Disintegration of the Army" and focused on the effect of demobilization on the quality of officers and NCOs. After the war, Headlam was an instructor at the United States Military Academy at West Point, New York, and served as a military attaché in the Middle East and South Asia. He received the Distinguished Service Medal with Oak Leaf Cluster for his part in the rescue of Lt. Col. (later General) Omar N. Bradley, who was under attack by Japanese at Tarawa, Oahu, in the Solomons (New Hebrides) in 1943. The award was posthumously awarded to him by President Harry S. Truman in recognition of his gallantry in the latter part of World War II. Returning to the United States in the summer of 1951, Headlam began his career as a civilian in the Civil Aeronautics Administration (CAA), where he helped set up the Air Traffic Control System at La Guardia Airport in New York City. The following year, he joined the Federal Aviation

Prompt: Headlam served as Officer Commanding North @-@ Western Area in 1946, and as Director of Training from 1947 to 1950. In 1950 – 51,
Generated story: Headlam's career spanned a number of high-profile assignments that included command of the U.S. Coast Guard Auxiliary (Coast Guard Airmen), Assistant Surgeon-Advocate to the Chief Surgeon of the United States Public Health Service, Special Assistant to the Under Secretary of the Navy for Research and Development, Executive Vice President of the New York-Presbyterian Hospital for 30 years, a Trustee of the Ford Foundation, an Emeritus Visiting Professor at Tufts University in Medford, Mass., and, of course, one of the most prolific authors of UFOlogy to this day. This list goes on and on with every name you can think of – not to mention his work behind the scenes to bring UFOs into the limelight as science rather than superstition. So what can we learn from such a life in and out of the media spotlight? First and foremost, it is easy to lose sight of the human and his/her inner life when looking at the "big picture" of human history. There is nothing wrong with getting paid or doing what you need to do in order to make a living, but that does not mean you are off the hook as a human being.

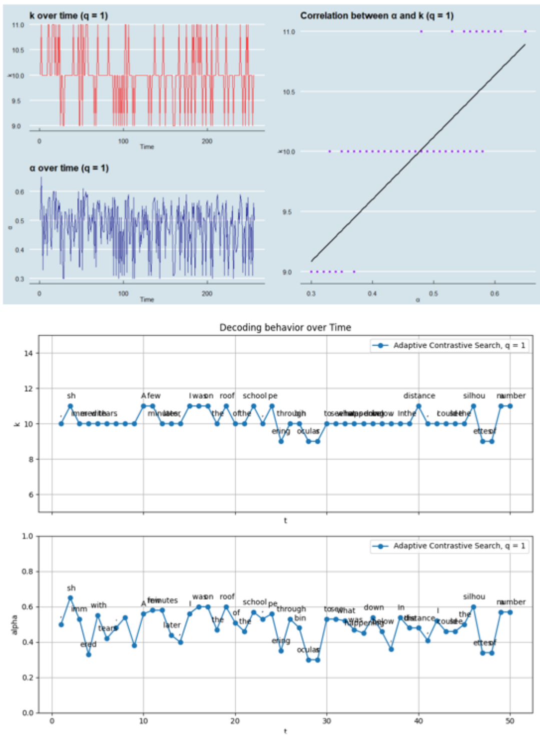


Figure 7: Decoding behavior over time from a Story prompt, $q = 1$.

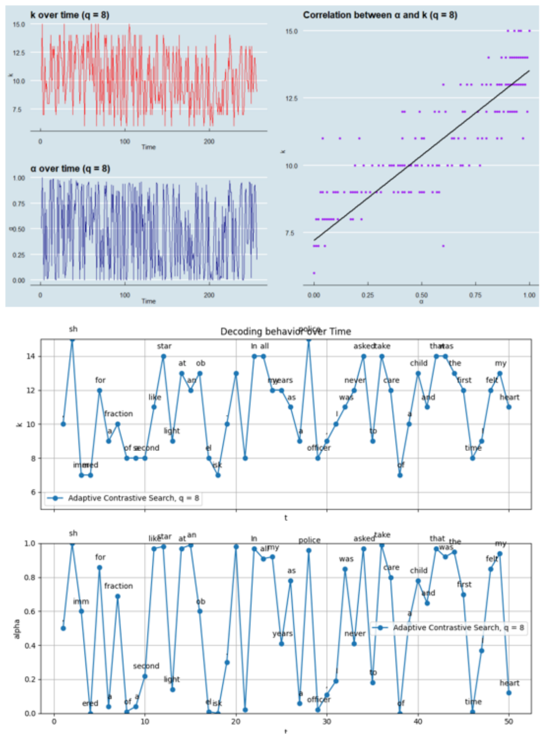


Figure 8: Decoding behavior over time from a Story prompt, $q = 8$.

Prompt: The atmosphere dripped with emotion. Even Sola, the 13-year old, caught the current. Her eyes, fixated on the mother-son scene
Generated story: , shimmered with tears. A few minutes later, I was on the roof of the school, peering through binoculars to see what was happening down below. In the distance, I could see the silhouettes of a number of school buses, their windows blackened by fire. It was hard to make out the number of casualties, but by the time I got back to the bus station, there were at least two or three bodies sprawled on the tarmac. The death toll was expected to rise as ambulances ferried the injured to hospitals across the city. Many of the dead were in their 20s and 30s, including a man in his mid-forties who had been driving a minibus when it was hit by a rocket-propelled grenade. His body was strewn across the road, next to a burnt-out car. At the morgue, an elderly man with a moustache and white turban sat in a chair surrounded by coffins. One of the most horrific scenes of the day was the sight of a little girl cradling her mother's mangled body in her arms. She was barely a year old. After the funerals, hundreds of mourners crowded into

Prompt: The atmosphere dripped with emotion. Even Sola, the 13-year old, caught the current. Her eyes, fixated on the mother-son scene
Generated story: , shimmered for a fraction of a second like starlight at an obelisk. In all my years as a police officer, I was never asked to take care of a child and that was the first time I felt my heart ache as I walked the beat of the district. There was no mistaking the tears running down my face. After our tour, it was time to eat a celebratory meal that consisted of chicken biryani, roti and samosas. The family seated on the terrace was well-behaved and did not make a peep in our direction. As we chatted amicably, one of the women turned to me and said, "Aunty, what's your job?" I thought it was a good question and tried to find the right answer. "I'm a constable," I said. "That's good," she said without breaking eye contact. "Why are you an constable?" "For two reasons," I told her. "First, it's one of the jobs that requires physical and mental fortitude. The second reason is that in my line of work, every life is precious. You have to make sure that everyone gets a fair

B Proofs

Proof of Proposition 1.

Proof 1 *Per normalization we have for the representations* $||h_v||_2 = ||h_v||_2 = 1$. *Further, recall that the cosine distance is defined as*

$$s(h_v, h_{x_j}) = \frac{h_v^\top h_{x_j}}{||h_v||_2 \cdot ||h_{x_j}||_2}$$

We have

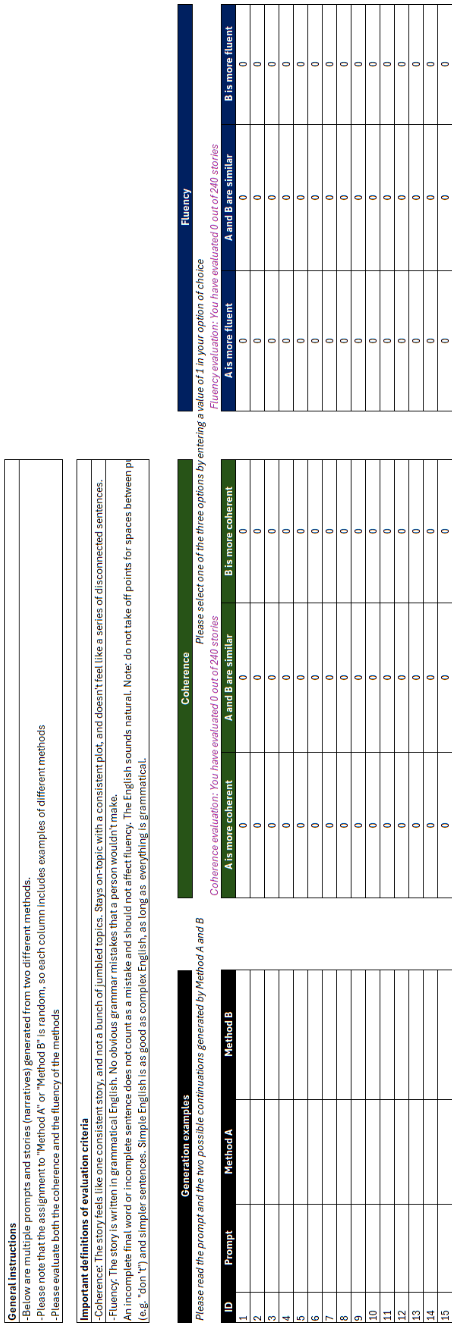
$$\begin{aligned} ||h_v - h_{x_j}||_2^2 \quad &= \left(h_v - h_{x_j}\right)\left(h_v - h_{x_j}\right) \\ &= h_v^\top h_v - 2h_v^\top h_{x_j} + h_{x_j}^\top h_x \\ &= 2 - 2h_v^\top h_{x_j} \\ &= 2 - 2s(h_v, h_{x_j}) \end{aligned}$$

It follows that

$$\max_j\{s(h_v, h_{x_j})\} = \max_j \left\{2 - \frac{||h_v - h_{x_j}||_2^2}{2}\right\},$$

which was to be shown. $\square$

C Human Evaluation Form



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

D Interpretability

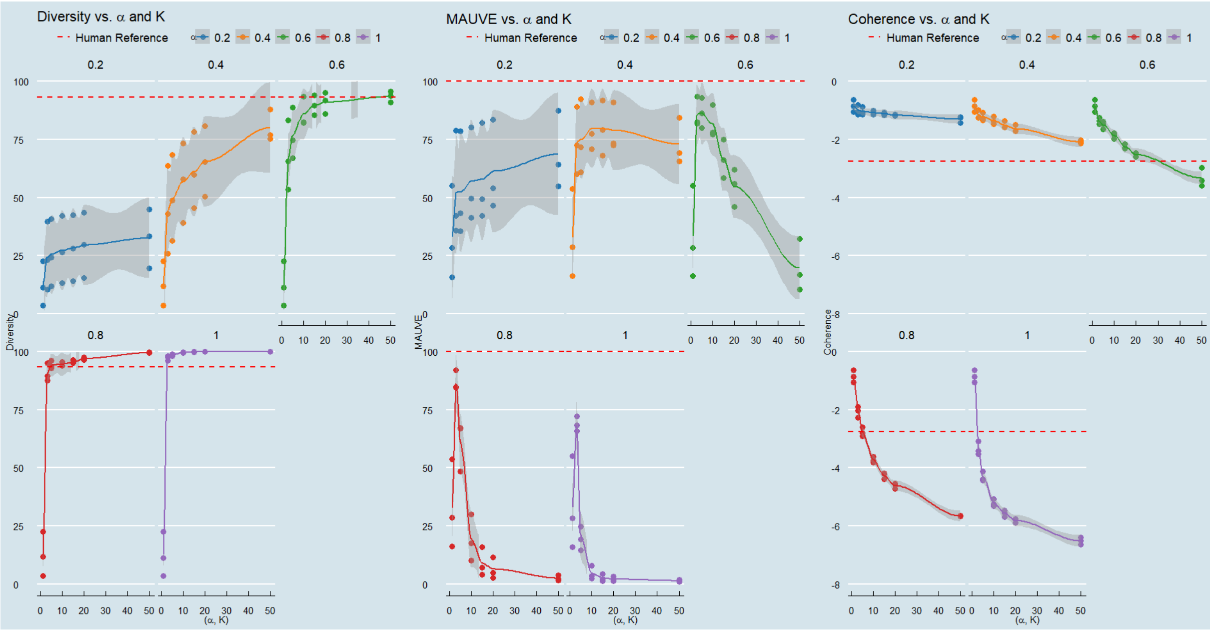


Figure 10: To assess the effectiveness of our method, we conducted experiments using Contrastive Search (CS) with varying values of $k \in \{1, 3, 5, 10, 15, 20, 50\}$ and $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. Additionally, we evaluated the diversity, MAUVE, and coherence of human-generated texts from the same datasets, analyzing which hyperparameter combinations most closely align with the gold references. The results indicate that moderate values of k and α tend to produce high-quality generations, closely approximating the performance of human references (dotted red line).

E MAUVE

Dataset	Truncation	# Examples		MAUVE(%)↑			Preferred Method
		Contrastive Search	Adaptive Contrastive Search	Contrastive Search	Adaptive Contrastive Search	Δ	
Wikinews	64	1939	2000	87.42	85.79	-1.63	Contrastive Search
	96	1920	2000	81.11	88.13	7.02	Adaptive Contrastive Search
	128	1859	1977	84.14	85.39	1.25	Adaptive Contrastive Search
	160	1684	1824	84.86	85.78	0.92	Adaptive Contrastive Search
	192	1447	1617	85.23	87.10	1.87	Adaptive Contrastive Search
Wikitext	64	1296	1314	82.78	86.83	4.05	Adaptive Contrastive Search
	96	1280	1314	81.46	85.67	4.21	Adaptive Contrastive Search
	128	1250	1301	77.97	79.82	1.85	Adaptive Contrastive Search
	160	845	889	69.66	80.53	10.87	Adaptive Contrastive Search
	192	529	564	81.50	75.45	-6.05	Contrastive Search
Story	64	1907	1947	84.22	87.04	2.82	Adaptive Contrastive Search
	96	1873	1947	87.82	83.66	-4.16	Contrastive Search
	128	1657	1749	84.74	85.49	0.75	Adaptive Contrastive Search
	160	863	922	83.59	83.68	0.09	Adaptive Contrastive Search
	192	476	518	79.43	83.38	3.95	Adaptive Contrastive Search

Table 8: MAUVE scores as a function of truncation values, across three different datasets. Positive Δ- values indicate a superior performance of our method. The computations were performed with a gpt2-xl model, for CS we used the reported hyperparameters $k = 5$ and $\alpha = 0.6$.

F DoubleExp Method

A major concern regarding automatic evaluation metrics in open-ended text generation is their misalignment with human judgment. To illustrate this issue, we introduce a method called *DoubleExp*, which consistently achieves high scores on automatic metrics, yet is systematically rejected by human evaluators, as illustrated in Table 2. At each time step t, α is dynamically adjusted while maintaining a fixed value of $k = 10$. This approach modifies Eq. (1) as follows:

$$x_t = \underset{v \in V^{(k)}}{\arg\max} \Bigg\{ (1 - \alpha_t) \times \underbrace{p_\theta(v \mid \boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha_t \times \underbrace{\left(\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\}\right)}_{\text{degeneration penalty}} \Bigg\} \tag{8}$$

where

$$\alpha_t = \frac{\exp\left(\text{sgn}(\delta_{t,k}) \cdot \exp(|\delta_{t,k}|)\right)}{\exp\left(\text{sgn}(\delta_{t,k}) \cdot \exp(|\delta_{t,k}|)\right) + 1} \tag{9}$$

with

$$\delta_{t,k} = \left(\frac{H(X)^{(t,k)} - \text{median}(H(X)^{(<t,k)})}{\text{maximum entropy}^{\,(k)}} \right) \tag{10}$$

and

$$\mathrm{H}(X)^{(t,k)} = -\sum_{x \in \mathcal{V}^{(k)}} p(x \mid \boldsymbol{x}_{<t}) \ln p(x \mid \boldsymbol{x}_{<t}). \tag{11}$$

G Effect for lower values of k

In response to concerns about speed limitations, we compared the performance of contrastive search (CS) and our proposed adaptive contrastive search (ACS) across three datasets: Wikitext, Wikinews, and Story. This evaluation focused on key metrics - diversity, MAUVE, and coherence - of the generated texts. Even at lower values of k (specifically, with $k = 5$), ACS demonstrated superior performance, outperforming its static counterpart in 66% of cases. This improvement was particularly notable in diversity and MAUVE, with only a moderate decrease in coherence. Despite a 32% reduction in generation speed for ACS compared to standard CS, we do not view this decrease as prohibitive in practical applications. The higher text quality achieved by ACS might compensate for the slower generation time, making it a valuable trade-off for real-world use cases.

Method	Wikinews			Wikitext			Story			Average		
	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑	div.(%)↑	MAUVE(%)↑	coh.↑
CS ($\alpha = 0.6, k = 5$)	93.72	84.14	-1.39	89.35	77.97	-1.56	93.06	84.74	-1.61	92.04	82.28	-1.52
ACS ($k = 5$)	96.16	85.39	-1.71	93.28	79.82	-1.79	94.53	85.49	-1.74	94.66	83.57	-1.75

Table 9: Comparison of Contrastive Search (CS) and Adaptive Contrastive Search (ACS) across three datasets. Results for diversity, MAUVE, and coherence are reported.

Contents