The Stable Entropy Hypothesis: An Entropy-Based Analysis of Neural Text Degeneration.

Anonymous authors
Paper under double-blind review

Abstract

State-of-the-art language generation models can degenerate under greedy decoding when applied to open-ended text generation problems such as text completion, story generation, or dialog modeling. This degeneration usually shows up as self-repetition or copying from the context. In this paper, we analyze this problem from a conditional entropy-centric perspective. We postulate that "human-like" generations usually lie in a narrow and nearly flat conditional entropy zone. Excessive lower- and upper-bound violations of this entropy zone correlate with degenerate and incoherent behavior, respectively. Our experiments show that this flat entropy zone exists across models, tasks, and domains and that the violation of this zone correlates with degeneration and incoherence. [1]

1 Introduction

Current state-of-the-start transformer-based (Vaswani et al., 2017) large language models have made a tremendous amount of progress on generation tasks such as summarization (Zhang et al., 2020; Lewis et al., 2020), machine translation (Raffel et al., 2022; Liu et al., 2020), and dialog generation (Roller et al., 2020; Shuster et al., 2022), story generation (Brown et al., 2020), etc. However, these models can degenerate under greedy decoding, especially when applied to open-ended text generation problems such as text completion, story generation, or long-form question answering. The degeneration shows up as self-repetition or copying from the context (Holtzman et al., 2019; Welleck et al., 2019).

This neural text degeneration problem (Holtzman et al., 2019) is mitigated by employing well-tuned stochastic decoding methods. These methods uniformly sample from either an annealed or a truncated distribution and are known to produce more coherent generations with less repetition that score high on generation quality metrics such as Mauve (Pillutla et al., 2021) and human acceptability judgments. Though these stochastic decoding methods generate fluent text, they are not without problems. These methods rely on random sampling at each time step and have been known to generate less contextual (Li et al., 2020), factual (Lee et al., 2022a), and verifiable (Massarelli et al., 2020) generations.

In this paper, we examine this degeneration conundrum — i.e., the unexpected degeneration of deterministic decoding methods in an open-ended generation setting and the surprising relative robustness of stochastic decoding methods, through the lens of entropy of the conditional distribution of the language model[2]. We start by presenting a finding that, under the context distribution from the human-generated data, the mean conditional entropy at time step t (computed across corpus) of a language model remains stable over the length of the generation. We refer to this mean conditional entropy as the *stable entropy baseline*, and a narrow band around the stable baseline as the *stable entropy zone*. Our experiments establish that the stable entropy phenomenon exists across the tasks, domains, and model combinations.

[1]We will open-source our code upon acceptance.

[2]For brevity, from hereon, we will refer to the entropy of the conditional distribution of the model either as entropy or conditional entropy based on the context.

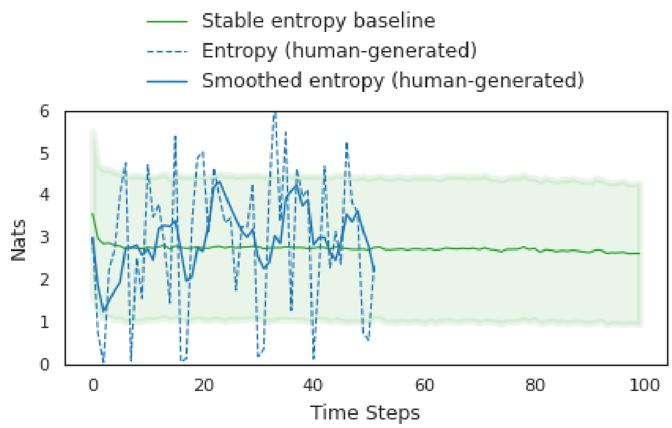


Figure 1: **The stable entropy zone annotated.** The faint green line is the entropy baseline computed under the human-generated data distribution. We refer to it as the **stable entropy baseline**. The green hue around it represents its $\alpha = 1.5$ standard deviation and is the **stable entropy zone**. The dashed and solid blue lines represent the conditional entropy and smoothed conditional entropy of single human completion from Appendix Table 4.

Our analysis shows that, in an open-ended generation setting, deterministic decoding algorithm's generations suffer a catastrophic drop in conditional entropy over the sequence length. In contrast, conditional entropy under well-tuned stochastic decoding algorithms remains mostly confined within the stable entropy zone. We use this finding to posit that any decoding algorithm whose resultant conditional entropy across time steps stays primarily within this narrow stable entropy zone will result in more coherent and less degenerate text. We refer to this hypothesis as the *stable entropy hypothesis* (SEH). We empirically validate this hypothesis by showing a strong correlation between generation quality and entropy zone violations in text completion setting. We show that the lower-bound entropy violations are strongly correlated with repetition, and upper-bound violations correlate with incoherence.

Next, in Section 5, we situate our work in the broader context of entropy-based analysis of language generation and analysis of degeneration phenomena in neural text generation. Finally, in Section 6, we conclude by summarizing our findings and discussing the potential broader application of this analysis, from designing a decoding algorithm that balances the degeneration and contextuality to training approaches that can leverage entropy violations as a negative reward for RL-style auxiliary loss.

Summarizing, in this paper, we analyze the degeneracy of deterministic decoding algorithms and the robustness of stochastic decoding algorithms through the lens of the mean conditional entropy of the model. We observe that, for greedy and beam search, the conditional entropy drops catastrophically over the sequence length, whereas, for well-tuned sampling-based methods, the conditional entropy mostly stays within a narrow flat range. We further introduce the notion of stable entropy baseline and stable entropy zone to analyze this behavior. We show that deterministic decoding's catastrophic entropy drop correlates with the degeneracy in open-ended generation settings. We also show that sampling-based methods in open-ended generation setups respect the stable entropy zone, and this adherence to the stable entropy zone strongly correlates with generation quality. We refer to this phenomenon as the stable entropy hypothesis and empirically validate it in the text completion setting.

2 Stable Entropy Analysis

In this section, we formalize the notion of the entropy baseline, the stable entropy baseline, and the stable entropy zone. We will then empirically demonstrate the existence of such a zone across combinations of language models, domains, and tasks.

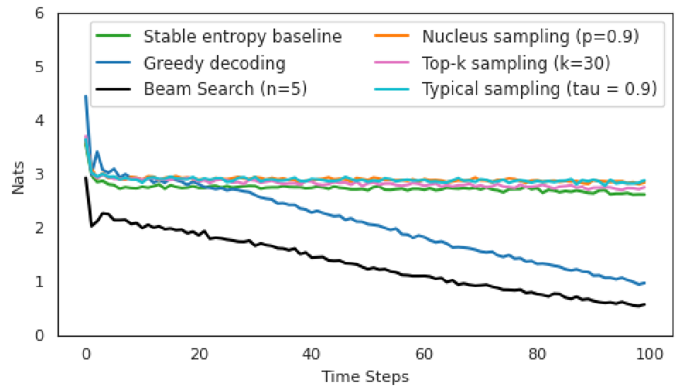


Figure 2: **Entropy baseline under various decoding algorithms.** We observe that the entropy baseline under greedy and beam search drops near-monotonically over the sequence length. Well-tuned sampling-based methods nearly follow the stable entropy baseline.

2.1 Entropy Baseline and Zone

Let p_θ be an autoregressive language model trained on a dataset $\mathcal{D} = \{(x_i, {}^i w_1^T)\}_{i=1}^N$, parameterized by θ. Given an input or source, x, and previous tokens or context till time step t, w_1^t, the conditional entropy of the model is defined as

$$\mathcal{H}(p_\theta, w_1^t; x) = \mathop{\mathbb{E}}_{w \sim p_\theta(\cdot|w_1^t)} - \log p_\theta(w|w_1^t; x) \tag{1}$$

We now define the **entropy baseline** as the mean conditional entropy at time step t under the context distribution induced by data $\boldsymbol{d}$ at time t, $w_1^t \in \boldsymbol{d}$:

$$\mu_{\mathcal{H}}(t; \boldsymbol{d}, p_\theta) = \mathbb{E}_{w_1^t \in textbfd} \big[\mathcal{H}(p_\theta, w_1^t)\big]. \tag{2}$$

The dataset $\boldsymbol{d}$ can either be generated by a sampling model p_θ using a decoding algorithm $\mathcal{A}$ or can be human-generated (i.e., $\mathcal{D}$).

We now define the **stable entropy baseline** as the mean entropy at time step t under the human-generated context distribution at time t, $w_1^t \in \mathcal{D}$:

$$\mu_{\mathcal{H}}(t; \mathcal{D}, p_\theta) = \mathbb{E}_{w_1^t \in \mathcal{D}} \big[\mathcal{H}(p_\theta, w_1^t)\big]. \tag{3}$$

Note, here $\mathcal{D}$ indicates human-generated data distribution, i.e., the stable entropy baseline is an entropy baseline computed under human-generated data distribution.

Figure 2 visualizes the stable entropy baseline, i.e., entropy baseline under human-generated data distribution, and entropy baselines under various decoding algorithms for GPT-2 XL model in the Wikipedia text completion setting [3]. We can observe that the entropy baseline under greedy and beam search drops near-monotonically over the sequence length. In contrast, the stable entropy baseline and the entropy baselines computed under the data generated using sampling-based methods such as top-k, nucleus, and typical sampling nearly follow the stable entropy baseline.

Next, we define the **stable entropy zone** as a zone around the stable entropy baseline that covers a major fraction of conditional entropy (across data points in the corpus) of the model under the human-generated data distribution. We define it in terms of the model's standard deviation. We choose 1.5 standard deviation ($\sigma_{\bar{\mathcal{H}}}(t; \mathcal{D}, p_\theta)$) around the stable

[3]Exact details on the setting are discussed in the section 2.2.1.

entropy baseline as the stable entropy zone for our analysis. This span covers approximately 87% of smoothed conditional entropies induced under human-generated data distribution.[4]

Section 1 shows the stable entropy baseline and the stable entropy zone computed in the Wikipedia text completion setting computed under the GPT2-XL model. The solid green line shows the stable entropy baseline. As the figure shows, the mean conditional entropy under the human-generated context distribution (i.e., the stable entropy baseline) remains flat except for the first few steps—hence justifying the moniker of the stable entropy baseline. The region around the stable entropy baseline, represented with a green hue, is the stable entropy zone. We can observe that the stable entropy zone is also stable and flat; i.e., the variance of the model entropy does not vary much across the generation length.

2.2 Empirical Study of Stability

In this section, we show that the stable entropy zone generalizes across models, domains, and tasks. We will also analyze the decoding algorithms at the individual generation level with respect to stable entropy zone to qualitatively demonstrate that deterministic decoding algorithms violate stable entropy zone bounds. On the other hand, well-tuned stochastic methods mostly fall within the stable entropy zone. We start this section by discussing our experimental setup in the next section. We then discuss the empirical evidence for the existence of the stable entropy zone across models, domains, and tasks in the section 2.2.2, and analyze the decoding algorithms w.r.t. of the stable entropy zone in the section 2.2.3.

2.2.1 Models and Data

For our text completion experiments, we use the GPT-2 XL (Radford et al., 2019) model and Wikipedia data. We follow a similar setup as Krishna et al. (2022); i.e., we chunk Wikipedia documents into individual paragraphs, and use the first 256 tokens as prefixes, and limit the generation length to 128 tokens.

To demonstrate the generalizability of the stable entropy zone, we use a combination of five tasks spanning six different datasets and five different models. These tasks are text completion, dialog generation, summarization, and story generation. For text completion analysis, we use two models, GPT2-XL (Brown et al., 2020) and OPT (1.3B) (Zhang et al., 2022) and three different datasets from three different domains: the Wikipedia dataset (Krishna et al., 2022), a fiction dataset, PG19 Rae et al. (2019), and a news dataset, CC News (Hamborg et al., 2017). We evaluate CNN-DM (Hermann et al., 2015) dataset with the BART (Lewis et al., 2020) and the Pegasus (Zhang et al., 2020) models for summarization experiments and the BlenderBot (1B) (Roller et al., 2020) model on the Blended Skill Talk (Smith et al., 2020) for dialog generation experiments. For story generation, we evaluate the WritingPrompts (Fan et al., 2018) dataset with the GPT-2 XL model.

2.2.2 Stable Entropy Zone Generalizes Across Tasks, Domains, and Models.

Figure 3 shows the stable entropy baselines and the stable entropy zones across a combination of different tasks, models, and domains. Again, we observe that, except for the first few steps, the stable entropy baseline remains almost always flat and that the stable entropy zone almost always forms a narrow and flat band around it.

2.2.3 Sequence-Level Stability Analysis

We can observe a similar phenomenon while visualizing the completions for a given single prefix under human completion and various decoding algorithms. To do this, we first compute the smoothed conditional entropy for each decoding algorithm.

We can observe in Figure 1 that the unsmoothed entropy of the model contains many sudden drops or peaks. This high variance is local and token-level and can be attributed

[4]In our experiments (not reported here), we found that other reasonable choices of the width of stable entropy zone do not impact our conclusions.

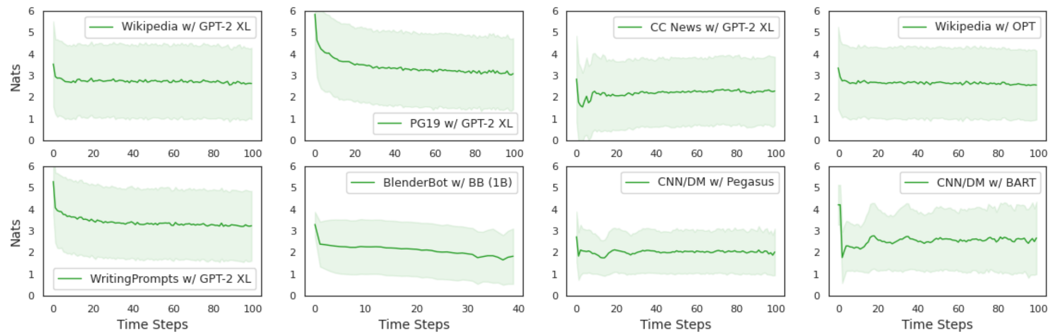


Figure 3: **Stable entropy baselines across models, tasks, and domains.** We observe that, except for the first few steps, the stable entropy baseline and the stable entropy zone are both nearly flat across the models (GPT2-XL, OPT, BlenderBot, Pegasus, and BART), tasks (text completion, story completion, dialog, and summarization), and domains (news, Wikipedia, and fiction).

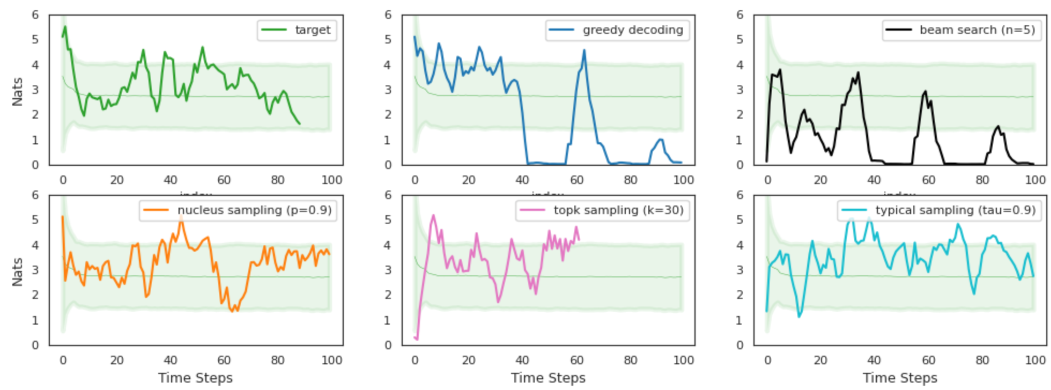


Figure 4: **Visualization of conditional entropy of a generation under various decoding algorithms.** Visualizing the smoothed conditional entropy for various decoding algorithms in a text completion setup given a prompt. We observe the catastrophic entropy drop in the case of the beam and greedy search. Stochastic algorithms try to stay in the stable entropy zone. Appendix Table 4 shows the prompt and generations corresponding to these visualizations.

to linguistic and tokenization phenomena such as collocations, the presence of function words, multi-token words, abbreviations, and punctuation in the sequence. This smoothing is necessitated to counter these confounds as our analysis is focused on how the entropy of the model evolves over the sequence and does not need to pay attention to token-level phenomena and variances.

We counter this token level confounds for our analysis by smoothing out the entropy. We compute smoothed entropy at time step t by averaging entropy over a small number ($U = 5$) of previous steps:

$$\bar{\mathcal{H}}(p_\theta, w_1^t) = 1/U \sum_{j=t-U}^{t} \mathcal{H}(p_\theta, w_1^j). \tag{4}$$

The solid blue line in Figure 1 represents this smoothed version of the same entropy represented by the dotted blue line. We apply similar smoothing for this analysis while computing the stable entropy baseline and the stable entropy zone.

Figure 4 visualizes the human-generated completions and completions generated by various decoding algorithms for a given single prefix in the Wikipedia text completion. The solid green line and solid green hue in the figures indicate the smoothed stable entropy baseline and the smoothed stable entropy zone. We use the 1.5 standard deviation of smoothed mean conditional entropy as the width of the stable entropy zone. We can clearly observe a catastrophic drop in smoothed conditional entropy for beam and greedy search, whereas the smoothed entropy of well-tuned sampling-based decoding algorithms stays mostly within the smoothed stable entropy zone. This phenomenon holds across the prefixes, and we show another example in Appendix A. These well-tuned stochastic decoding algorithms are also known to produce better completions (Holtermann et al., 2022) that are more coherent, less repetitive, and rate high on human acceptability judgments. We postulate that these two things might be related and try to quantify this correlation in the next section.

3 The Stable Entropy Hypothesis

Stochastic decoding algorithms that mimic the stable entropy baseline behavior also produce better completions (Holtermann et al., 2022). We hypothesize that decoding algorithms whose generation's smoothed conditional entropy stays mostly enclosed within the smoothed stable entropy zone will produce higher quality, coherent, and less repetitive text. We refer to this hypothesis as the **stable entropy hypothesis**.

Next, we empirically verify the stable entropy hypothesis and answer the following question:

Are violations of the stable entropy zone correlated with automatic measures of generation quality in more open-ended generation settings?

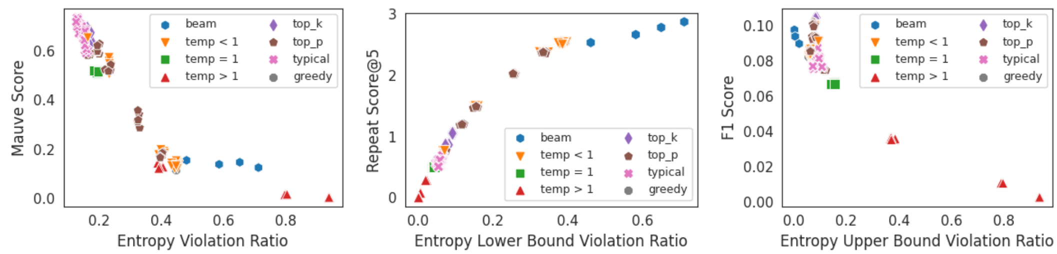
We do so by measuring the correlation between automatic metrics of text generation quality and entropy violation measures, defined as how often the smoothed conditional entropy of the generation under the choice of decoding algorithm violates the smoothed stable entropy zone. We continue to use the 1.5 standard deviation of the smoothed mean entropy as the width of the stable entropy zone.

3.1 Models, Data, and Metrics

We answer the question in the same text completion setting as discussed in Section 2.2; i.e., we use the GPT-2 XL (Radford et al., 2019) model and Wikipedia data from Krishna et al. (2022). In this setting, we evaluate various configurations of well-known decoding algorithms, namely, top-k sampling (Holtzman et al., 2019), nucleus sampling (Fan et al., 2018), temperature sampling, and typical decoding (Meister et al., 2023). See Appendix Section B for the configurations.

We use three automatic metrics to evaluate the performance of various decoding algorithms. **F1** computes the overlap between the generation and the "true" human-generated completion of the prefix, indicating whether the text is on-topic and contextually appropriate.[5] **Repeat@5** cumulatively measures the repetition across 1- to 5-grams weighted exponentially and normalized by length. A higher Repeat@5 indicates that the generation was more repetitive and dull. We discuss the exact computation of Repeat@5 in the Appendix Section C. **Mauve** (Pillutla et al., 2021), an automatic generation quality metric, evaluates generation quality in the open-ended generation setting and was shown to have a strong correlation with human acceptability judgments.

We measure entropy zone violations using three metrics. **Entropy lower-bound violation ratio (ELVR)** measures the ratio of instances of time steps when entropy falls below the lower bound of the stable entropy zone. **Entropy upper-bound violation ratio (EUVR)** measures the ratio of instances of time steps where entropy is larger than the upper bound of the stable entropy zone. The third metric, **entropy violation ratio (EVR)**, is the sum of the two ratios and measures the instances when entropy falls outside either the lower or the upper bound, and hence measures the overall adherence to the stable entropy zone.



(a) EVR vs Mauve Score (b) ELVR vs Repeat@5 (c) EUVR vs F1 Score

Figure 5: **Entropy violations vs repetition vs generation quality vs coherence.** Figure (a) shows that the Mauve score, a proxy for generation quality, correlates negatively ($\rho = -0.92$) with the entropy violations. Figure (b) shows lower entropy violations are strongly correlated ($\rho = 0.96$) with the repetition issue. Finally, Figure (c) shows that decodings schemes that result in high entropy produce relatively more incoherent text ($\rho = -0.93$).

Sampling Method	**F1**	**Rep. Score@5**	**3-gram rep.**	**Mauve**	**EVR**	**EUVR**	**ELVR**
Greedy	0.082	2.542	45.338	0.114	0.447	0.0560	0.391
Beam (n=5)	0.094	2.664	48.138	0.138	0.585	0.004	0.581
+3-gram block	0.102	0.666	0.063	0.476	0.170	0.014	0.155
Temperature Sampling							
$t = 0.5$	0.100	1.499	16.159	0.537	0.238	0.078	0.160
$t = 0.8$	0.091	0.761	3.146	0.653	0.162	0.093	0.069
$t = 1$	0.068	0.511	1.015	0.507	0.193	0.155	0.038
$t = 1.2$	0.035	0.287	0.178	0.130	0.403	0.383	0.020
Top-k Sampling							
$k = 30$	0.094	0.709	2.416	0.665	0.148	0.083	0.065
$k = 50$	0.091	0.666	2.016	0.667	0.144	0.083	0.062
Nucleus Sampling							
$p = 0.95$	0.075	0.557	1.289	0.592	0.169	0.122	0.047
$p = 0.9$	0.082	0.620	1.701	0.642	0.150	0.094	0.056
Typical Sampling							
$\tau = 0.2$	0.076	0.507	0.819	0.697	0.129	0.074	0.054
$\tau = 0.9$	0.082	0.615	1.725	0.622	0.154	0.093	0.061
Human completions	1.000	0.605	1.381	1.000	0.136	0.0631	0.0731

Table 1: **Quantitiative results for text completion analysis.** F1 score between the human-generated and model-generated completion measures the contextuality of the generations. 3-gram repeats measure the extent of repetition problem with the generations. Entropy Lower-Bound Violation Ratio (ELVR), Entropy Upper-Bound Violation Ratio (EUVR), and Entropy Violation Ratio (EVR) measure the frequency with which entropy lower-bound, entropy upper-bound, and both combined are violated.

3.2 Results

We present the correlation results in the text completion setting in Figure 5. We observe that Mauve scores have a strong negative correlation ($\rho = -0.92$) with the entropy violation ratio (EVR). This indicates that the decoding algorithm with more instances of smoothed conditional entropy falling outside the stable entropy zone usually has worse generation quality. We also observe a strong positive correlation ($\rho = 0.96$) between the Repeat@5 and the entropy lower-bound violation ratio (ELVR). This matches our observation that deterministic decoding methods that are prone to repetition and copying exhibit a catastrophic

[5]We filter out stop words from the sequences before computing F1 scores to ensure that these commonly occurring words do not confound contextuality judgment.

Algorithm 1 Entropy-Aware Sampling

Input: input x, model p_θ
Hyperparams: sampling $\mathcal{S}$, margin α, ngreedy g
Constants: SEM $\mu_\mathcal{D}$, SEV $\sigma_\mathcal{D}$
Initialize $n \leftarrow 0$
while $t \leq T$ **do**
  $p_t = p_\theta(y_{t-1}, x)$
  $w_t = \text{argmax}(p_t)$
  $\mathcal{H}_t = \text{Entropy}(p_t)$
  if $t > g$ & $\mathcal{H}_t > \mu_\mathcal{D} + \alpha \times \sigma_\mathcal{D}$ **then**
    $w_t = \text{Sample}(p_t, \mathcal{S})$
  end if
  $y_{t-1} = y_{t-1} w_t$
end while

drop in smoothed entropy, resulting in them falling below the stable entropy zone's lower bound. Finally, we observe a negative correlation ($\rho = -0.93$) between the entropy upper-bound violation ratio and F1 scores, indicating decoding methods with high conditional entropy (e.g., sampling with $t = 1.5$) usually produce a less coherent text.

This correlational analysis indicates that the model with fewer overall entropy violations results in higher-quality text and that the repetition problem is linked explicitly to lower entropy violations. We explain the second finding by hypothesizing that at some point, the decoding-induced context distribution diverges so much from the human-generated data distribution that the model decides to resort to copying or repeating. At this point, entropy drastically drops as the model knows exactly the next token it needs to generate. We present a few additional correlation plots in Appendix Section E.

Table 1 quantitatively verifies this hypothesis by showing that generations under greedy decoding and beam search degenerate as indicated by low Mauve score and high Repeat@5 and 3-gram repeats. This degeneration correlates with a high overall entropy violation ratio (EVR), a significant portion of which are entropy lower bound violations. High entropy upper bound violations, as is the case with sampling with a higher temperature hyperparameter ($t = 1.2$), indicate incoherence that can be attributed to a high amount of randomness, as suggested by very low Mauve and F1 scores. Furthermore, fewer entropy violations (both upper and lower bound), as in the case of top-k, nucleus, and typical sampling, as well as fewer repetitions, reasonable F1 score, and a high Mauve score, suggesting a correlation between better generation quality and entropy violations.

As observed in Figure 4, well-tuned sampling-based decoding algorithms mostly stay enclosed within the stable entropy zone. We present some qualitative examples in Appendix Table 4 of a prefix and completions, which show that the generations produced under sampling-based methods do indeed appear more coherent and less repetitive.

4 Entropy-Aware Sampling

In the previous section, we discussed how well-tuned stochastic decoding methods can alleviate degeneration issues in open-ended generation settings and how this improvement in generation quality also correlates with fewer stable entropy zone violations. These stochastic methods, though, rely on random sampling at each time step and have been noted to result in generation being less contextual (See Figure 5c and Table 1), more factually inaccurate Lee et al. (2022b) and less verifiable Massarelli et al. (2020).

In this section, we explore the possibility of leveraging the stable entropy analysis to propose a decoding algorithm that can better balance the trade-off between coherence and contextuality, and can overcome both, the deterministic decodings' degeneration and uniform randomness of the stochastic decoding algorithms. We hope the proposed algorithm can

Decoding Method	F1	Repeat@5	Mauve	EVR	Det%
Greedy	0.082	2.542	0.114	0.447	100
Beam (n=5)	0.094	2.664	0.138	0.585	100
+3-gram block	0.102	0.666	0.476	0.170	100
Typical Sampling ($\tau = 0.2$)	0.076	0.507	0.697	0.129	0
Top-k ($k = 30$)	0.094	0.709	0.665	0.148	0
Entropy-Aware Sampling (ours)					
$\tau = 0.2$, $\alpha = 0.5$, $g = 5$	0.088	0.70	0.690	0.140	58.7
$k = 30$, $\alpha = 0.5$, $g = 5$	0.101	1.06	0.657	0.178	59.45
Target completions	1.000	0.605	1.000	0.136	-

Table 2: **Entropy-Aware Decoding Text Completion Experiment.** We observe that entropy-aware decoding is competitive with typical sampling, the best performing stochastic decoding method from Table 1, on generation quality and repetitions while having higher F1 score indicating more contextually appropriate completions.

produce coherent, less repetitive, and less degenerate text despite acting greedily most of the time. We discuss one such possible algorithm below.

We motivate our proposed approach by highlighting that under low to moderate conditional entropy scenarios, the model is fairly confident in its prediction. We treat this as a proxy of the context appearing to be "in-distribution" to the context seen during training. In these scenarios, greedy decoding should suffice. When the conditional entropy of the model is high, as indicated by the upper-bound violations of the stable entropy zone, the model is less certain about its prediction. In such cases, chances of misprediction are high; i.e., the most probable token might not be the "correct" token. Hence, at these timesteps, i.e., when the model's conditional entropy is high and it breaches the upper bound of the stable entropy zone, we resort to sampling from the conditional distribution. While sampling, we can rely on any of the off-the-shelf sampling methods (denoted by $\mathcal{S}$ in Algorithm 1) such as top-k, nucleus, or typical sampling.

The proposed entropy-aware sampling (EAS) is outlined in Algorithm 1. From our stable entropy analysis, we know that the stable entropy baseline and stable entropy zone are both flat, and hence can be approximated using constants the stable entropy mean (SEM), $\mu_{\mathcal{D}}$, and the stable entropy variance (SEV), $\sigma_{\mathcal{D}}$ respectively. We also use a width coefficient α to control the width of the stable entropy zone. During generation, for the first g steps, we simply do greedy decoding. After that, if the model's entropy is above the upper bound of the stable entropy zone (i.e., $\mathcal{H}_t \geq \mu_{\mathcal{D}} + \alpha \times \sigma_{\mathcal{D}}$), we sample from the model distribution using the sampling algorithm $\mathcal{S}$.

4.1 Experiments

4.1.1 Model and Data

We benchmark entropy-aware decoding in the text completion setting. We use a similar setup and metrics as Section 3.1 for our text completion experiments. Additionally, we also report **%Det**, the percentage of the time entropy-aware sampling and other algorithms act deterministically.

4.1.2 Results

Table 2 presents the results for text completion experiments. We can observe that the entropy-aware decoding (with $\alpha = 0.5$, and typical sampling with $\tau = 0.2$) generates more on-topic and contextually appropriate, less repetitive, and higher quality text as indicated by high F1 score, low Repeat@5, and high Mauve score respectively. Also, the entropy-aware decoding method has a low entropy violation ratio supporting our hypothesis that this improved generation quality might be due to entropy-aware decoding's ability to stay within the stable entropy zone. Entropy-aware sampling acts greedily most of the time (nearly 60%) as indicated by Det% measure.

5 Discussion and Related Work

Entropy-based Decoding Approaches: Recently, a few stochastic methods have been proposed that use entropy or related concepts to truncate the probability distribution. **Typical decoding** (Meister et al., 2023) induces sparsity by selecting a subset of tokens whose likelihood is closest to the entropy of the model. The number of tokens is controlled by the cumulative probability we want to retain in the distribution. **Mirostat decoding** (Basu et al., 2021) modifies top-k sampling where the k is dynamic and controlled in such a way that it ensures that the generation has similar perplexity to the human-generated data. Recently proposed η**-sampling** (Hewitt et al., 2022) samples from the tokens whose probability is greater than η, which is defined as a function of the model's entropy. All these decoding methods are fully stochastic, sampling at each time step, introducing uniform randomness, which might hurt the contextuality and the factuality (Lee et al., 2022a) and verifiability (Massarelli et al., 2020) of the generation. In contrast, the stable entropy analysis allows for the design of algorithms that only act stochastically when the upper bound of the stable entropy zone is violated. This behavior would hopefully result in a higher F1 score, indicating more on-topic and contextual generations. We leave the design of such an algorithm for future work.

Stable Entropy Hypothesis and Uniform Information Density Hypothesis: Uniform information density (UID) hypothesis (Jaeger & Levy, 2006) states that subject to the grammar constraint, humans prefer sentences that distribute information, measured in terms of surprisal, equally across the linguistic signal (Meister et al., 2020).

The UID hypothesis differs from the stable entropy hypothesis in some crucial aspects. First, UID is concerned about the distribution of surprisal in human communication due to bandwidth and efficiency constraints. The stable entropy hypothesis explicitly deals with the behavior of decoding algorithms in the neural language generation context and how "human-like" generations can be achieved by ensuring that the model's entropy stays within a narrow zone around the stable entropy baseline. Second, UID is defined in terms of surprisal which takes into account the token generated/uttered at each time step. In contrast, the stable entropy hypothesis is defined in terms of conditional entropy over time t, which is the expected surprisal over vocabulary under the model distribution. Third, the stable entropy hypothesis is more accommodating as it just expects the decoding algorithm's conditional entropy to fall within the stable entropy zone, whereas the UID hypothesis expects the model's generation's surprisal to be flat or stable for it to be more "human-like". Given the differences, we plot surprisal for the same prefix as in Figure 4 in Appendix Figure 8. Similar to the catastrophic drop in entropy under greedy and beam search, we observe that text generated under greedy and beam search do not follow the UID hypothesis either and suffer a similar drop in surprisal.

Stable Entropy Hypothesis, Expected Information Hypothesis, and Local Typicality: The Expected Information Hypothesis(EIH), proposed by Meister et al. (2022), formally states that text perceived as human-like typically encodes an amount of information close to the expected information content of natural language strings, i.e., in the interval $-log p(y) \in [H(p) - \epsilon, H(p) + \epsilon]$ for a natural language. Text that falls outside of this region is likely perceived as unnatural. This differs from the stable entropy hypothesis in three important respects. First, the EIH, unlike the stable entropy hypothesis (SEH), deals with entropy at the sequence level, hence missing the temporal component of SEH. Additionally, like the UID hypothesis, it analyzes the information content, whereas the stable entropy hypothesis analysis is based on the conditional entropy of the model. Thirdly, the anchor entropy value in the case of the EIH is computed under ancestral sampling. This differs from SEH, where the anchor value, the stable entropy baseline, is computed on human-generated data.

Meister et al. (2023) further extended this idea to add a temporal component and define a related concept of local typicality. Local typicality states that the information content of every word in natural-sounding sentences must be close to the expected information content under p, i.e., the conditional entropy given prior context (Meister et al., 2023).

Similar to EIH, local typicality bounds the surprisal or the information content. The stable entropy hypothesis, in contrast, bounds the entropy of the conditional distribution of the model. Second, the stable entropy zone is anchored around the stable entropy baseline that is defined in terms of the conditional entropy of the model under human-generated context distribution, whereas the local typicality uses the conditional entropy of the model under the distribution induced by the current decoding algorithm. Thus, this definition cannot be used to analyze the decoding algorithms' behaviors. A case in point is the analysis of degenerate behavior under deterministic decoding in an open-ended generation setting. In this setting, the anchor value—i.e., the entropy of the model under greedy decoding, will itself drop catastrophically, resulting in surprisal always staying within the bounds, indicating that strings generated under greedy decoding satisfy local typicality and hence are natural sounding.

6 Conclusion

In this paper, we presented the stable entropy hypothesis that states that the entropy of natural language stays in a narrow zone around the stable baseline, which is defined as the mean conditional entropy of the model under the huaman-generated context distribution. We verify this hypothesis in the text completion setting, showing that fewer violations of the stable entropy zone correlate with fewer repetitions and higher generation quality. We posit that the stable entropy hypothesis and the related concept of the stable entropy zone can be leveraged in two significant ways. First, it can be used to design either backtracking algorithms similar to Look-back Decoding Xu et al. (2023) or a selectively sampling algorithm that only samples with upper-bound of the stable entropy zone is breached. We hypothesize that the mostly deterministic nature of either of these decoding schemes will also improve the factuality of the generation, an important problem that needs to be solved before the wide-scale deployment of large language models. We leave this analysis for future work. Additionally, SEH analysis can also be used to design auxiliary RL-style objectives that enforce the stable entropy hypothesis, resulting in more robust language generation models that do not degenerate under deterministic decoding algorithms.

References

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. Mirostat: A Neural Text Decoding Algorithm that Directly Controls Perplexity, January 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, June 2020.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pp. 218–223, March 2017. doi: 10.5281/zenodo.4120316.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching Machines to Read and Comprehend. *arXiv:1506.03340 [cs]*, June 2015.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation Sampling as Language Model Desmoothing, October 2022.

Carolin Holtermann, Anne Lauscher, and Simone Ponzetto. Fair and argumentative language modeling for computational argumentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7841–7861, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.541. URL https://aclanthology.org/2022.acl-long.541.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]*, April 2019.

T. Jaeger and Roger Levy. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. RankGen: Improving Text Generation with Large Ranking Models, May 2022.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Factuality Enhanced Language Models for Open-Ended Text Generation, June 2022a.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Factuality Enhanced Language Models for Open-Ended Text Generation, June 2022b.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y.-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training. *arXiv:1911.03860 [cs]*, May 2020.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation, January 2020.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. How Decoding Strategies Affect the Verifiability of Generated Text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 223–235, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.22. URL https://www.aclweb.org/anthology/2020.findings-emnlp.22.

Clara Meister, Ryan Cotterell, and Tim Vieira. If beam search is the answer, what was the question? In *EMNLP 2020*, pp. 2173–2185, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.170.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. On the probability–quality paradox in language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 36–45, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.5. URL https://aclanthology.org/2022.acl-short.5.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, January 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00536.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers, November 2021.

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

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive Transformers for Long-Range Sequence Modelling. November 2019. doi: 10.48550/arXiv.1911.05507.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y.-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. *arXiv:2004.13637 [cs]*, April 2020.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y.-Lan Boureau, Melanie Kambadur, and Jason Weston. BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage, August 2022.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL https://aclanthology.org/2020.acl-main.183.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural Text Generation with Unlikelihood Training. *arXiv:1908.04319 [cs, stat]*, August 2019.

Nan Xu, Chunting Zhou, Asli Celikyilmaz, and Xuezhe Ma. Look-back decoding for open-ended text generation. *arXiv preprint arXiv:2305.13477*, 2023.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11328–11339. PMLR, November 2020.

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. OPT: Open Pre-trained Transformer Language Models, June 2022.

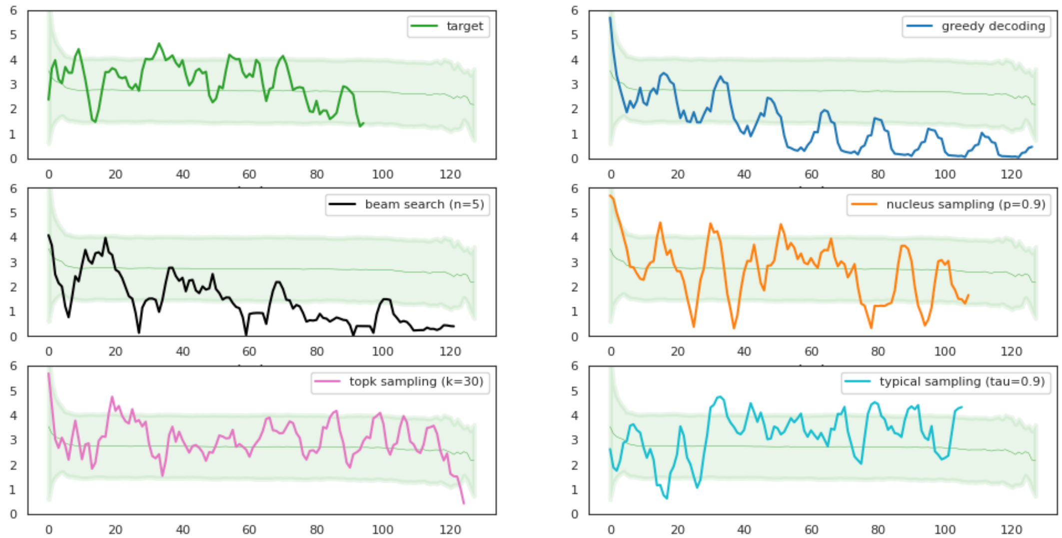


Figure 6: **Visualization of conditional entropy of a generation under various decoding algorithms.** Visualizing the smoothed conditional entropy for various decoding algorithms in a text completion setup given a prompt. We observe the catastrophic entropy drop in the case of the beam and greedy search. Stochastic algorithms try to stay in the stable entropy zone. Appendix Table 4 shows the prompt and generations corresponding to these visualizations.

A Visualization of entropy of a generation under various decoding algorithms

B Various Configurations of Decoding Algorithm Evaluated in Section 3.1

We evaluate the following configurations of stochastic decoding algorithms for the stable entropy hypothesis experiments. We run each algorithm on three different seeds.

- Top-K Sampling (k): 5, 10, 30, 50, 100,
- Nucleus Sampling (p): 0.15, 0.25, 0.4, 0.5, 0.75, 0.9, 0.95,
- Ancestral Sampling with Temperature (t): 0.001, 0.01, 0.1, 0.2, 0.5, 0.8, 1.0, 1.2, 1.5, 3.0,
- Typical Sampling (τ): 0.2, 0.25, 0.5, 0.75, 0.9, 0.95.

C Repeat@5

We modify the Repeat@5 metric proposed in **?**, to capture the repetition at various n-gram levels. The modified metric is just a length-normalized version of the original metric. We compute Repeat@5 as

$$\text{Repeat@5} = \log_2\left(\frac{\sum_{i=1}^{5} 2^i \times \#\,\text{i-grams}}{\#\,\text{cuml n-grams}}\right) \times \#\,\text{1-grams}/\#\textit{Tokens} \tag{5}$$

where $\#\,\text{cuml n-grams} = \sum_{i=1}^{5} \#\,\text{i-grams}$.

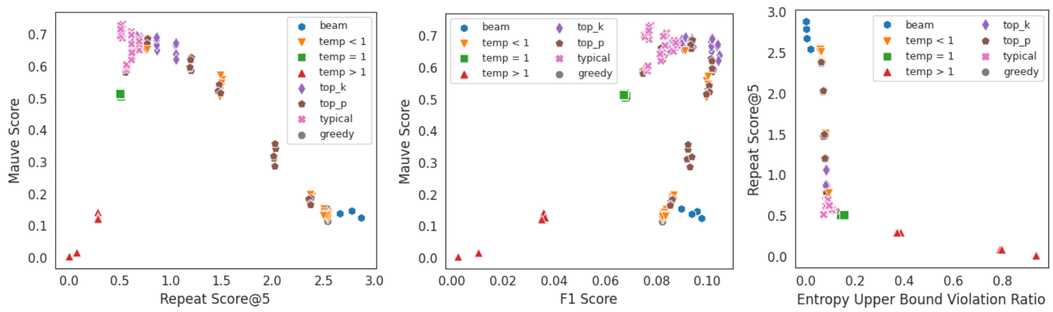
Intuitively, the metric captures average numbers of repetitions per token while exponentially weighing 1-gram to 5-gram repetitions.

D Qualitative Examples

Context: ...With the coming of the Interstate Highway System in Michigan, the MSHD planned to convert the eastern section of US 2 to a freeway between St. Ignace and Sault Ste. Marie. In planning maps from 1947, this highway corridor was included in the system that later became the Interstates.
Greedy: The western section of US 2 was renumbered as M-54 in 1949. The eastern section of US 2 was renumbered as M-28 in 1950. The western section of US 2 was renumbered as M-54 in 1951. The eastern section of US 2 was renumbered as M-28 in 1952. ...
BeamSearch (k=5): In the 1950s and 1960s, the M-54 designation was used to designate the section of US 2 between Sault Ste. Marie and Sault Ste. Marie. In the 1970s, the M-54 designation was used to designate the section of US 2 between Sault Ste. Marie and Sault Ste. Marie. In the 1980s, the M-54 ...
Top-k Sampling (k=30): In 1948, the western end of US 2 was rerouted to the eastern point near the old US 2/US 3 junction in Sault Ste. Marie. A section of the original M-1 was restored to US 2 at Wakefield before the first interstate was completed. ...
Typical Sampling (τ=0.9): Michigan Department of Transportation letters to Congress, dated 18 November 1947 and 3 June 1948, approved the abandonment of the business US 2 segment of US 2 in favor of an expanded US 12. Approximately 3.5 miles of US 12 was completed between Ann Arbor and Sault Ste. Marie between 1964 and 1965. ...

Table 3: Generation examples using various decoding methods in a text completion setting using GPT-2 XL model. Greedy and beam search results in catastrophic degeneration (repetitions highlighted in red) whereas stochastic methods generate relatively more coherent completions.

E Additional Correlational Plots And Configurations for Stable Entropy Hypothesis Analysis



(a) Repeat@5 vs Mauve Score (b) F1 Score vs Mauve Score (c) Repeat@5 vs EUVR.

Figure 7: **Additional Correlation Plots for Text Completion Experiments.** Figure (a) shows too many repeats (beam and greedy search, and temperature sampling ($T << 1$)) and too few repeats (for temperature sampling ($T >> 1$)) both hurt generation quality. Figure (b) shows that, among the stochastic decoding methods, top-k sampling balances the contextuality and generation quality conundrum the best. Finally, Figure (c) shows a strong negative correlation between the repetition issue and entropy upper zone violations indicating that mostly lower-bound violations are mostly responsible for copying and repetitions.

F Surprisal Under Different Decoding Algorithms

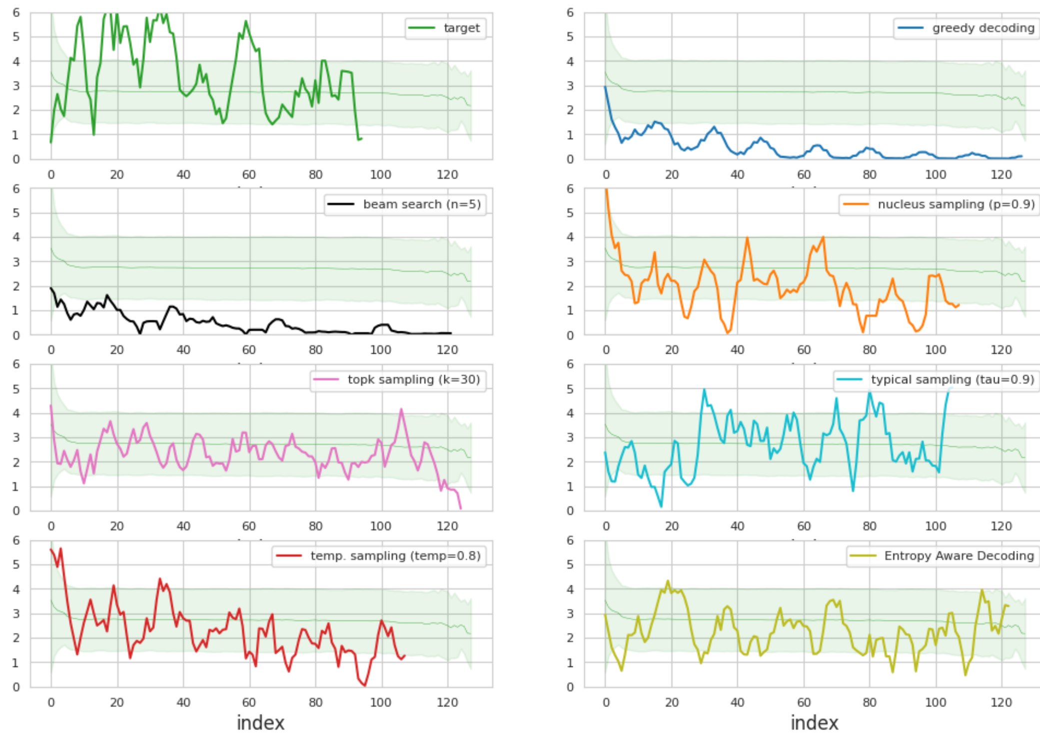


Figure 8: **Visualization of surprisal of various decoding algorithms.** Visualizing the smoothed surprisal (smoothing window size 5) for various decoding algorithms in a text completion setup for the prompt from Table 4. The faint green line in the background is the smoothed stable entropy baseline and is used to represent the target information rate. We observe the catastrophic drop in surprisal for beam and greedy search. Stochastic algorithms oscillate near the target information rate.

Context: ...With the coming of the Interstate Highway System in Michigan, the MSHD planned to convert the eastern section of US 2 to a freeway between St. Ignace and Sault Ste. Marie. In planning maps from 1947, this highway corridor was included in the system that later became the Interstates.	
Greedy: The western section of US 2 was renumbered as M-54 in 1949. The eastern section of US 2 was renumbered as M-28 in 1950. The western section of US 2 was renumbered as M-54 in 1951. The eastern section of US 2 was renumbered as M-28 in 1952. ...	
BeamSearch (k=5): In the 1950s and 1960s, the M-54 designation was used to designate the section of US 2 between Sault Ste. Marie and Sault Ste. Marie. In the 1970s, the M-54 designation was used to designate the section of US 2 between Sault Ste. Marie and Sault Ste. Marie. In the 1980s, the M-54 ...	
Top-k Sampling (k=30): In 1948, the western end of US 2 was rerouted to the eastern point near the old US 2/US 3 junction in Sault Ste. Marie. A section of the original M-1 was restored to US 2 at Wakefield before the first interstate was completed. ...	
Typical Sampling (τ=0.9): Michigan Department of Transportation letters to Congress, dated 18 November 1947 and 3 June 1948, approved the abandonment of the business US 2 segment of US 2 in favor of an expanded US 12. Approximately 3.5 miles of US 12 was completed between Ann Arbor and Sault Ste. Marie between 1964 and 1965. ...	
Entropy-Aware Decoding: ... The western section of US 2 was renumbered as M-47 in December 1949. With the interchange in Epoufette opened in 1949, the state route continued to serve businesses in M-48. A traffic jam of 3,000 cars per hour formed on US 2 in July 1950 when three women got into a disagreement on the roadway. After about an hour, the cars were cleared and the issue resolved. ...	

Table 4: Generation examples using various decoding methods in a text completion setting using GPT-2 XL model. Greedy and beam search results in catastrophic degeneration (repetitions highlighted in red) whereas stochastic methods generate relatively more coherent completions.

Person 1: I do not like the taste, texture or thought of eating meat.
Person 2: So it has nothing to do with whether or not eating meat is ethical? There is a lot of debate around the ethics of eating meat.
Person 1:

Beam Search	I don't like the taste, texture, or thought of eating meat. I don't like the taste, texture, or thought of eating meat.
EAD	It's always good to hear different opinions. What would you say is the reason you don't eat meat?

Person 1: Sounds like you had a great childhood then!
Person 2: You could say that. What was your childhood like?
Person 1: I can't really remember a lot, but I did play a lot of basketball with my friends
Person 2: I love basketball! I took part in many basketball tournaments and played in high school as well. What position do you play?
Person 1: I still practice basketball regularly, not a particular position thought
Person 2:

Beam Search	I played a lot of basketball in high school as well. I was a point guard. What position do you play?
EAD	I still play as well. I really enjoy it. I'm good at it. I usually play with my friends.

Person 1: We want to have at least 6 kids. We are from Italy, and we love having big families!
Person 2: Oh, that is awesome! Do you and your wife come from a big family?
Person 1: My wife does, I am an only child. My mom is in the medical field and did not really have time for kids.
Person 2: I'm an only child as well. I wish that I had siblings growing up but it did allow me special moments with my parents.
Person 1:

Beam Search	My wife and I want to have at least 6 kids. We are from Italy and want to have a big family.
EAD	I always wanted siblings, but my parents did not want me to have any.

Table 5: Dialog qualitative examples where beam search produces at least two 3-gram repeats.