The Stable Entropy Hypothesis and Entropy-Aware Decoding:
An Analysis and Algorithm for Robust Natural Language Generation

Kushal Arora [1] **Timothy J. O'Donnell** [1] **Doina Precup** [1,2] **Jason Weston** [3] **Jackie C.K. Cheung** [1]

Abstract

State-of-the-art language generation models can degenerate when applied to open-ended generation problems such as text completion, story generation, or dialog modeling. This degeneration usually shows up in the form of incoherence, lack of vocabulary diversity, and self-repetition or copying from the context. In this paper, we postulate that "human-like" generations usually lie in a narrow and nearly flat entropy band, and violation of these entropy bounds correlates with degenerate behavior. Our experiments show that this stable narrow entropy zone exists across models, tasks, and domains and confirm the hypothesis that violations of this zone correlate with degeneration. We then use this insight to propose an entropy-aware decoding algorithm that respects these entropy bounds resulting in less degenerate, more contextual, and "human-like" language generation in open-ended text generation settings. [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 both strongly conditioned generation tasks such as summarization (Zhang et al., 2020; Lewis et al., 2020) and machine translation (Raffel et al., 2022; Liu et al., 2020) and more open-ended generation tasks such as dialog generation (Roller et al., 2020; Shuster et al., 2022), story generation (Brown et al., 2020), etc. Almost all these large-scale language models are trained by maximizing the log-likelihood of the training sequences. This, one would

assume, will result in likelihood-maximizing decoding algorithms such as greedy and beam search, producing outputs that match the informativeness, coherence, and quality of generation of the training data. This assumption holds for more strongly conditioned tasks but, for more open-ended generation tasks, deterministic decoding methods produce repetitive and dull outputs, referred to as degeneration in Holtzman et al. (2019). In these open-ended generation settings, stochastic decoding methods can help. 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.

In this paper, we examine this degeneration conundrum — i.e., the degeneration of deterministic decoding methods in an open-ended generation setting while being robust for strongly conditioned generation tasks, and the relative robustness of well-tuned stochastic decoding methods in open-ended generation setups, 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 training data, the mean entropy of a language model remains stable over the length of the generation. We refer to this mean entropy as the *stable entropy baseline*, and a narrow band around the stable baseline spanned as the *stable entropy zone*. In our experiments, we establish that the stable entropy phenomenon exists across the tasks, domains, and model combinations.

In our analysis, we observe that, in an open-ended generation setting, deterministic decoding algorithms suffer a catastrophic drop in entropy over the sequence length. In contrast, 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 entropy across timesteps stays mostly 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

---

[1]Mila & McGill University [2]DeepMind [3]Meta AI. Correspondence to: Kushal Arora <kushal.arora@mail.mcgill.ca>.

[1]Implementation of Entropy Aware Decoding: https://github.com/kushalarora/transformers/blob/main/src/transformers/generation_utils.py#L1894.

Source code for the stable entropy hypothesis analysis and the entropy-aware decoding experiments: https://github.com/kushalarora/stable_entropy_hypothesis

---

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

between generation quality and entropy zone violations.

Next, we attempt to explain the relative robustness of deterministic decoding methods such as beam search for strongly grounded tasks using the stable entropy analysis. The stable entropy hypothesis would suggest that for strongly conditioned tasks, we will not observe a similar catastrophic drop in entropy under greedy decoding. Our experiments confirm this observation; i.e., for strongly conditioned generation tasks, the smoothed entropy under beam search mostly respects the stable entropy zone bounds.

Finally, we leverage the stable entropy analysis to propose a new entropy-aware decoding method that can avoid degeneration while acting greedily most of the time. On two tasks: text completion and dialogue generation, we show that entropy-aware decoding results in a less degenerate, more contextually appropriate, and "human-like" generation.

Summarizing, in this paper, we analyze the disparate behavior of various decoding algorithms through an entropy-aware perspective. We show that deterministic decoding's catastrophic drop in entropy may explain their degeneracy in open-ended generation settings and that this catastrophic drop is absent in strongly grounded tasks indicating their relative robustness in tasks such as machine translation and summarization. We also show that sampling-based methods in open-ended generation setups do respect the stable entropy hypothesis and this adherence to the stable entropy hypothesis strongly correlates with generation quality. Finally, we leverage the insights from our analysis to propose a novel entropy-aware decoding method that results in a more "human-like" generation while acting greedily most of the time.

2. Stable Entropy Analysis

2.1. Stable Entropy Zone

Let p_θ be an autoregressive language model trained on a dataset $\mathcal{D}$, parameterized by θ. Given an input or source, x, and previously generated tokens or context, w_1^t, the entropy of the model is defined as

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

The entropy of the model can suffer from high variance (See Figure 1). This variance can possibly be attributed to linguistic and tokenization phenomena such as collocations, the presence of function words, multi-token words, abbreviations, and punctuation in the sequence. To reduce the variance for our analysis, we smooth out the entropy. We compute smoothed entropy at time step t by averaging
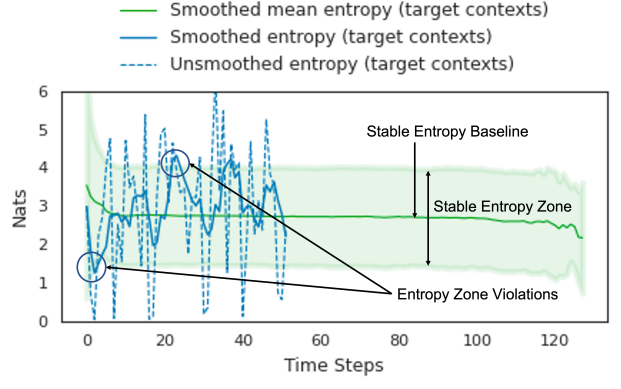

Figure 1. **The Stable Entropy Zone annotated.** The dashed and solid blue lines represent the entropy and smoothed entropy of single target completion. The faint green line is the mean smoothed entropy computed under the target context distribution. We refer to it as the **stable entropy baseline**. The green hue around it represents its 1.5 standard deviation and is the **stable entropy zone**. Breaches of the stable entropy zone's upper and lower bound are referred to as **entropy upper bound violations (EUV)** and **entropy lower bound violations (ELV)** respectively and combined they are called **entropy bound violations (EV)**.

entropy over a small number (U) of previous steps[3]:

$$\bar{\mathcal{H}}(p_\theta, w_1^t) = 1/U \sum_{j=t-U}^{t} \mathcal{H}(p_\theta, w_1^j). \quad (2)$$

We now define the **stable entropy baseline** as the mean smoothed entropy at timestep t under the target context distribution at time t, $w_1^t \in \mathcal{D}$:

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

Next, we define the **stable entropy zone** as a zone around the stable entropy baseline that covers a major fraction of smoothed entropy (across data points in the corpus) of the model under the target 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 entropy baseline as the stable entropy zone for our analysis. This span covers approximately 87% of smoothed conditional entropies induced under target distribution.[4]

2.1.1. EMPIRICAL STUDY OF STABILITY

In this section, we show that the smoothed mean entropy of a model under the target context distribution remains stable; i.e., it remains nearly flat—hence justifying the moniker of

---

[3]We drop the input x from equations for brevity.

[4]Other choices of the width of the stable entropy zone are possible. In our experiments, we found that similar choices of the width of stable entropy zone do not impact our conclusions.
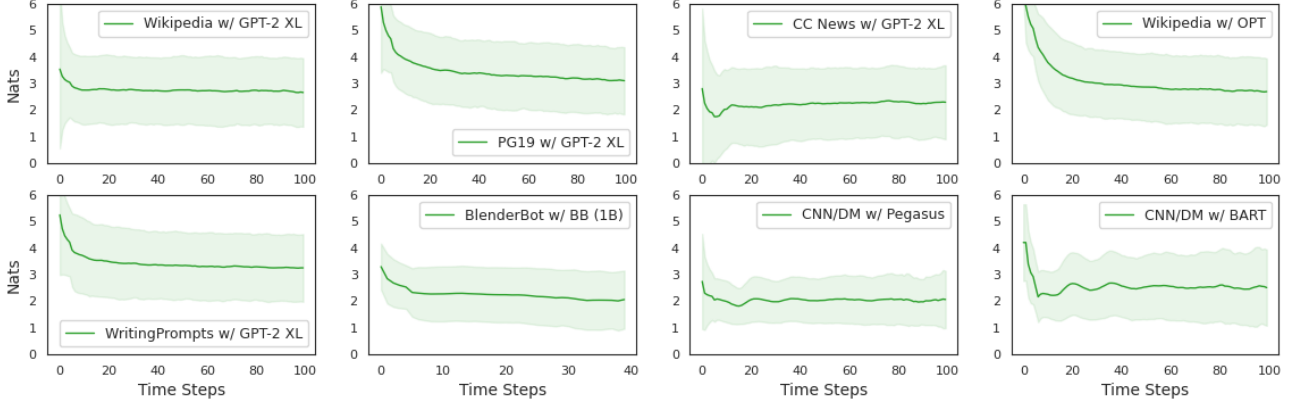
Figure 2. **Stable entropy baselines across models, tasks, and domains.** We observe that the stable entropy baseline is nearly flat and the stable entropy zone is narrow across models (GPT2-XL, OPT, BlenderBot, Pegasus, and BART), tasks (text completion, story completion, dialog, and summarization), and domains (news, Wikipedia, and fiction).

the stable entropy baseline. We also show that the stable entropy zone—i.e., the area spanned by a fraction of its standard deviation—is narrow and mostly flat. We start by demonstrating this phenomenon in a text completion setup, then show that the stable entropy zone generalizes across models, domains, and tasks.

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 (Brown et al., 2020) model.

2.1.2. RESULTS

Stable Entropy Zone Does Exist.   Figure 1 shows the stable entropy baseline and the stable entropy zone computed

Setup	MSE	Slope	Intercept
GPT2-XL/Wiki (TC)	0.010	-0.0028	2.88
GPT2-XL/PG19 (TC)	0.090	-0.009	3.91
OPT/Wiki (TC)	0.207	-0.0130	3.80
BB1/BST (D)	0.011	-0.0143	2.64
BART/CNN-DM (S)	0.069	0.0071	2.74

Table 1. **Can stable entropy baseline be modeled using a flat 1-D line?**. Modeling smoothed mean entropy using a 1-D line yields a very small mean squared error and a near-zero slope. This indicates a flat 1-D line can model the stable entropy baseline.

in the Wikipedia text completion setting. The blue dotted line shows the entropy of the model under the target context distribution for a single prefix (See Appendix Table 6 for the prefix and the target completion). We can observe in Figure 1 that the unsmoothed entropy of the model contains many sudden drops or peaks thus necessitating the need to smooth it out for analysis. The solid blue line represents this smoothed entropy under the target context distribution. The solid green line shows the stable entropy baseline. As the figure shows, the mean smoothed entropy under the target context distribution (i.e., the stable entropy baseline) remains nearly flat except for the first few steps. 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 narrow (≈ 2 Nats) and flat.

Stable Entropy Zone Generalizes Across Tasks, Domains, and Models.   Figure 2 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.

We further quantify this observation by finding the line of best fit for the smoothed mean entropy (Table 1). We observe very low mean squared error loss and near-zero slope coefficients indicating that a flat 1-D line can fit smoothed entropy under the target context distribution with very few outliers. The variance in intercepts indicates that the stable entropy baseline must be evaluated individually for each model and dataset combination.

2.2. Stable Entropy Hypothesis

Figure 3 visualizes the completions for a given single prefix from Appendix Table 6 for various decoding algorithms in the Wikipedia text completion. We can clearly observe a catastrophic drop in smoothed entropy for beam and greedy search whereas smoothed entropy of well-tuned sampling-based decoding algorithms stays mostly within the stable entropy zone. These stochastic decoding algorithms are also known to produce better completions (Holtermann et al., 2022). We postulate that these two things might be related.

We hypothesize that decoding algorithms whose generation's smoothed entropy stays mostly enclosed within the stable entropy zone will produce higher quality, coherent, less repetitive, and more "human-like" text. We refer to this hypothesis as the **stable entropy hypothesis**.

Next, we empirically verify the stable entropy hypothesis by measuring the correlation between automatic metrics of text generation quality and entropy violation measures.

2.2.1. EXPERIMENTS

In this section, we answer the following two questions:

- Are violations of the stable entropy zone correlated with automatic measures of generation quality in more open-ended generation settings?
- Does the stable entropy hypothesis also hold for more strongly conditioned tasks where deterministic search strategies do not degenerate?

Models, Data, and Metrics: We answer the first question in the same text completion setting as discussed in Section 2.1.1; 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 F.1 for the configurations.

We use four automatic metrics to evaluate the performance of various decoding algorithms. **F1** computes the overlap between the generation and the "true" completion of the prefix, indicating whether the text is on-topic and contextually appropriate.[5] **Repeat Score@5** cumulatively measures the repetition across 1- to 5-grams weighted exponentially and normalized by length. [6] **3-gram repeats** measures the number of 3-gram repeats in the generated sequence. A higher Repeat Score@5 and 3-gram repeats indicate that the generation was more repetitive and dull. **Mauve** (Pillutla et al., 2021), a recently introduced 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 when smoothed entropy falls below the lower bound of the stable entropy zone. Similarly, **entropy upper-bound violation ratio (EUVR)** measures the ratio of instances where smoothed 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 ratio of instances when entropy falls outside either the lower or the upper bound.

To answer the second question, we contrast the performance of beam search with a fixed beam size ($n = 5$) between strongly conditioned tasks and more open-ended generation tasks. We use summarization and machine translation as our prototypical strongly conditioned language generation tasks and text completion and dialog generation as open-ended generation tasks. For summarization experiments, we report the ROUGE-1 score on two datasets and model combinations. We evaluate the CNN-DM dataset with the Pegasus (Zhang et al., 2020) model and the Arxiv (Cohan et al., 2018) dataset with the BigBird-Pegasus (Zaheer et al., 2020) model. We report BLEU scores for our machine translation experiment on the WMT 2017 dataset (de-en split) (Bojar et al., 2017) with two models, M-BART (Tang et al., 2021) and Opus (Tiedemann & Thottingal, 2020). We contrast the performance of beam search on these tasks with two more open-ended generation tasks namely, text completion and dialog generation. We report the F1 score for dialog generation on the Blended Skills Talk dataset (Smith et al., 2020) with two different sizes of the BlenderBot model (Roller et al., 2020) (90M and 1B parameters). For text completion experiments, we report results on the Wikipedia dataset from Krishna et al. (2022) with GPT2-XL (Radford et al., 2019) and OPT (1.3B) (Zhang et al., 2022) models. Except for text completion models, all the other models are encoder-decoder models.

---

[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.

[6]We discuss the exact computation of Repeat Score@5 in the Appendix Section G.
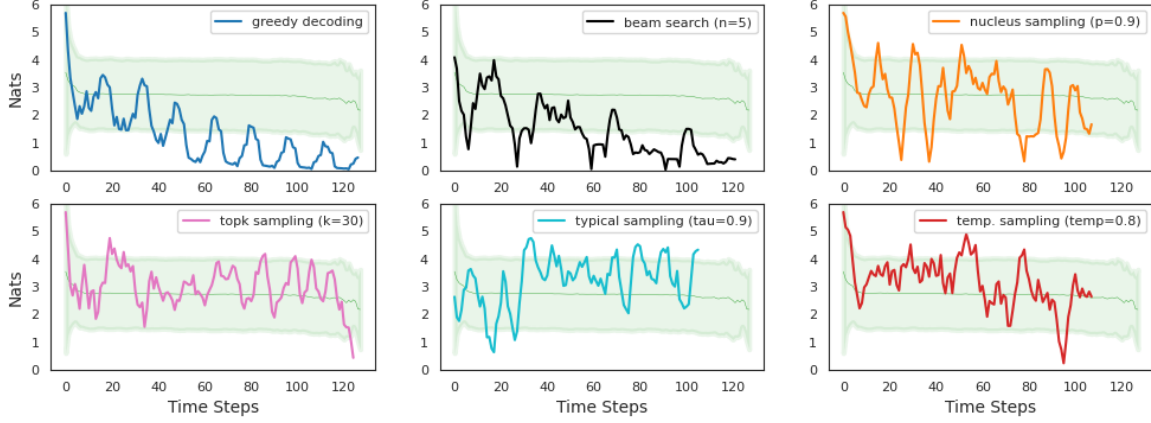
Figure 3. **Visualization of entropy of various decoding algorithms.** Visualizing the smoothed entropy for various decoding algorithms in a text completion setup given a prompt. We observe the catastrophic entropy drop in the case of the beam and greedy search. Stochastic algorithms try to stay in the stable entropy zone. Table 6 shows the prompt and generations corresponding to these visualizations.
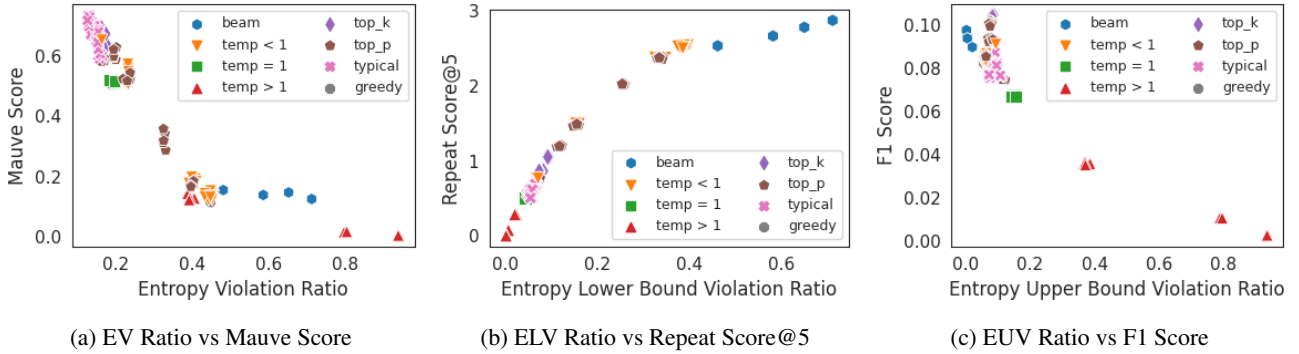

(a) EV Ratio vs Mauve Score

(b) ELV Ratio vs Repeat Score@5

(c) EUV Ratio vs F1 Score

Figure 4. **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 ($\rho = -0.93$).

2.2.2. RESULTS

Stable Entropy Hypothesis Holds for Text Completion.
We present the correlation results in the text completion setting in Figure 4. We observe that Mauve scores have a strong negative correlation ($\rho = -0.92$) with the entropy violation ratio (EVR). This indicates a decoding algorithm that generates more instances of smoothed 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 Score@5 and the entropy lower-bound violation ratio (ELVR). This matches our observation that deterministic decoding methods which are prone to repetition and copying exhibit a catastrophic 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 entropy (e.g., sampling with $t = 1.5$)

usually produce a less coherent text. We present additional correlation plots in Appendix Section B.

Appendix Table 5 quantitatively verifies this hypothesis by showing that generations under greedy decoding and beam search degenerate as indicated by low Mauve score and high Repeat Score@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. And, fewer entropy violations (both upper and lower bound), as in the case of top-k, nucleus, and typical sampling, do correlate in fewer repetitions, reasonable F1 score, and a high Mauve score, suggesting better generation quality under these decoding schemes.

Figure 3 visualizes the completions from various decoding algorithms for a single prefix. We can observe that well-tuned sampling-based decoding algorithms mostly stay enclosed within the stable entropy zone. The prefix and completions used to generate these visualizations are presented in Appendix Table 6 and they qualitatively show that the generations produced under sampling-based methods do indeed appear more coherent and less repetitive.

When Beam Search Does Not Degenerate. Table 2 presents the generation results decoded using beam search ($n = 5$) for both more open-ended (rows 1-5) and strongly conditioned generation settings (rows 6-10). For text completion tasks, we observe a catastrophic drop in entropy and poor generation quality as indicated by high entropy lower-bound violation ratio (ELVR) and low Mauve scores respectively. In contrast, for the strongly-conditioned generation tasks (last 5 rows), the beam search performs substantially better as indicated by near state-of-the-art ROUGE-1 and BLEU scores for summarization and machine translation respectively. Also, we observe far fewer entropy violations for these strongly-conditioned tasks indicating that they mostly remain within the stable entropy zone, thus respecting the stable entropy hypothesis. Dialog generation, an open-ended generation task modeled using an encoder-decoder model falls squarely between the two and has relatively high ELVR scores. A larger and more performant (in terms of F1 score) dialog model (BB-1B) does produce fewer entropy violations.

3. Entropy Aware Decoding

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 lower stable entropy zone violations. These stochastic methods, though, rely on uniform random sampling at each time step, which might results in generation being less contextual and more factually inaccurate (Lee et al., 2022). In this section, we use the insights from the stable entropy analysis to propose a decoding algorithm that can overcome the degeneration issues of deterministic algorithms while still acting greedily most of the time thus avoiding uniform randomness introduced by the stochastic decoding algorithms. The goal is that such an algorithm would result in a coherent and more on-topic and contextually appropriate generation.

The proposed entropy-aware decoding (EAD) method is outlined in Algorithm 1. Entropy-aware decoding lets the model decode greedily most of the time and intervenes in two scenarios: 1.) when the entropy of the model breaches the upper bound, and 2.) when the entropy of the model breaches the lower bound N consecutive times. We re-

---

Algorithm 1 Entropy-Aware Decoding

Input: input x, model p_θ
HP: sampling $\mathcal{S}$, patience N, margin α, ngreedy g
Initialize $n \leftarrow 0$
while $t \leq T$ **do**
    $p_t = p_\theta(y_{t-1}, x)$
    $w_t = \text{argmax}(p_t)$
    if $t \leq g$ **then**
        continue
    end if
    $\mathcal{H}_t = \text{Entropy}(p_t)$
    if AboveStableEntropyZone($\mathcal{H}_t, \alpha$) **then**
        $w_t = \text{Sample}(p_t, \mathcal{S})$
    end if
    if BelowStableEntropyZone($\mathcal{H}_t, \alpha$) **then**
        $n = n + 1$
    else
        $n \leftarrow 0$
    end if
    if $n > N$ **then**
        $y_{t-1}, p_t = \text{BackOffTo}(t - N)$
        $t = t - N$
        $w_t = \text{NextRankedToken}(p_t)$
        $n \leftarrow 0$
    end if
    $y_{t-1} = y_{t-1}w_t$
end while

---

fer to these interventions as **Entropy Upper-Bound Interventions (EUI)** and **Entropy Lower-Bound Interventions (ELI)** respectively.

Upper-bound violation of the stable entropy zone indicates that the model is less certain about its prediction. In such scenarios, chances of miscalibration are high; i.e., the most probable token might not be the "correct" token. Hence, in the scenario where entropy 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 methods (denoted by $\mathcal{S}$ in Algorithm 1) such as top-k, top-p, or typical sampling.

For lower-bound violations, we wait until N consecutive violations, as entropy drop at any time-step might be due to the presence of multi-token words, abbreviations, or other tokenization quirks. If the entropy is below the lower bound of the entropy zone for N consecutive steps, we back off N steps to the index when it was last above the threshold. At that index, we ignore the current most likely token and select the next highest-ranked token. We continue executing the backoff strategy until we select a token that does not lead N consecutive steps of entropy lower-bound violations.

Setup	Task	Mauve (↑)	F1 (↑)	BLEU (↑)	ROUGE-1 (↑)	ELVR (↓)
GPT2-XL/Wiki	Text Completion	0.137				0.588
GPT2-XL/PG19	Text Completion	0.048				0.690
OPT/Wiki	Text Completion	0.141				0.605
BB-90M/BST	Dialog		0.110			0.300
BB-1B/BST	Dialog		0.153			0.185
Pegasus/CNN-DM	Summarization				43.37 (-4.6)	0.121
BigBird/Arxiv	Summarization				46.05 (-4.9)	0.147
Opus/WMT	Machine Translation			35.04 (-6.0)		0.084
MBART/WMT	Machine Translation			37.90 (-3.2)		0.101

Table 2. **Beam Search does not degenerate for strongly conditioned generation tasks.**. We observe that for conditional generation tasks such as summarization, and machine translation, beam search performs well as indicated by high ROUGE-1 and BLEU scores respectively. The number in the bracket indicates the drop in performance compared to the state of the art. These tasks and model combinations also have few entropy zone violations. For open-ended tasks such as text competition, the beam search catastrophically degenerates resulting in a poor Mauve score and high entropy violation ratio. See Appendix F for model details and state-of-the-art baselines.

3.1. Experiments

3.1.1. MODEL AND DATA

We benchmark entropy-aware decoding on two open-ended generation tasks: text completion and dialog generation.

Text Completion We use a similar setup and metrics as Section 2.2.1 for our text completion experiments. Additionally, we also report **%Det**, the percentage of the time entropy-aware decoding and other algorithms act deterministically and **#Backoff**, an average number of times entropy-aware decoding algorithm resorted to backoffs.

Dialog Generation For dialog generation experiments, we use the 90M parameter BlenderBot model (Roller et al., 2020) and report results on the Blended Skills Talk dataset (Smith et al., 2020). We flatten the dialogs in the dataset by concatenating the previous utterances in a dialog. This forms the context for generating the next utterance. We limit the size of the context to 80 words and only keep the latest utterances that fit within the context size. We limit the maximum length of a generated utterance to 128 tokens. We report our dialog generation results on three metrics, namely, F1, Repeat Score@5, and entropy violation ratio. For these experiments, we follow the standard practice and do not remove the stop words from the target and the generated utterances while computing the F1 score.

3.1.2. RESULTS

Text Completion Results: Table 3 presents the results for text completion experiments. We can observe that the entropy-aware decoding (with patience window, $N = 5$, margin $\alpha = 0.8$, 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 Score@5 and 3-gram repeats, and high Mauve score respectively. Also, the entropy-aware decoding method has the lowest 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 decoding acts greedily most of the time (nearly 60%) as indicated by Det% measure.

Dialog Generation Results: Table 4 presents the result for our dialog generation experiments. We observe stochastic decoding methods do reduce repetition but at the cost of a lower F1 score. This reduction in the F1 score can be attributed to uniform randomness introduced by stochastic decoding methods. Entropy-Aware Decoding (patience window, $N = 5$, margin $\alpha = 0.25$), with both top-k and top-p sampling, successfully reduces the repetition issue while achieving the highest F1 score.

Human Evaluation We also performed human evaluation comparing 100 samples the entropy-aware decoding with typical sampling ($\tau = 0.2$) to typical sampling ($\tau = 0.2$) and 100 samples of entropy-aware decoding with top-k sampling (k=30) with top-k sampling (k=30). We asked 8 human evaluators to compare 25 samples each, collecting two annotations per sample in the process. EAD ($\tau = 0.2$) was preferred over typical sampling 57% of the time whereas EAD ($k = 30$) was preferred over top-k sampling 64% of the time.

Decoding Method	F1	Rep. Score@5	Mauve	EVR	Det%	#Backoffs
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 Decoding (ours)						
$\tau = 0.2, N = 5, \alpha = 0.5, g = 10$	0.090	0.696	0.688	0.116	58.8	1.97
w/o ELI	0.092	0.773	0.652	0.155	58.7	0
$k = 30, N = 5, \alpha = 0.5, g = 5$	0.100	0.941	0.683	0.135	59.6	2.88
w/o ELI	0.101	1.06	0.657	0.178	59.45	0
EAD w/o EUI $N = 5, \alpha = 0.5$	0.089	2.124	0.232	0.308	100	8.90
Target completions	1.000	0.605	1.000	0.136	-	-

Table 3. **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 5, on generation quality and repetitions while having higher F1 score indicating more contextually appropriate completions.

Decoding Method	F1	Rep. Score@5	EVR	Det%
Greedy	0.115	1.229	0.162	100
Beam (n=5)	0.118	1.171	0.305	100
Top-k (k=30)	0.112	0.489	0.155	0
Nucleus ($p = 0.9$)	0.116	**0.526**	0.160	0
EAD ($k = 30$)	0.125	0.674	0.130	62
w/o ELI	**0.125**	0.731	0.155	64
EAD ($p = 0.9$)	**0.125**	0.685	**0.130**	62
w/o EUI	**0.126**	0.742	0.156	64
EAD w/o EUI	0.114	1.166	0.136	100

Table 4. **Entropy-Aware Decoding Dialog Generation Experiments.** We observe that entropy-aware decoding produces the highest F1 score among all the methods irrespective of the choice of sampling algorithm. It achieves this while reducing the repetitions encountered when generating with greedy or beam search.

4. Discussion and Related Work

Entropy-based Becoding 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 target 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 of the generation (Lee et al., 2022). In contrast, entropy-aware decoding only samples if the upper bound of the stable entropy zone is violated. This behavior results in a higher F1 score indicating more on-topic and contextual generations.

Stable Entropy Hypothesis and Uniform Information Density Hypothesis: Uniform information density (UID) hypothesis (Levy, 2005; 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 is related to the stable entropy hypothesis as both predict the "stable" behavior of the model's prediction under human context distribution. But they also differ in some crucial aspects. First, the stable entropy hypothesis is defined in the terms of entropy, which is expected surprisal over vocabulary under the model distribution. Second, the stable entropy hypothesis is more accommodating as it just expects the model's entropy to fall within a narrow zone whereas the UID hypothesis expects the model's generation's surprisal to be nearly flat or stable for it to "human-like". We plot surprisal for the same prefix as in Fig-

ure 3 in Appendix Figure 9. Similar to the catastrophic drop in entropy under greedy and beam search, we observe that greedy and beam search do not follow the UID hypothesis and suffer a similar drop in surprisal.

Stable Entropy Hypothesis and Expected Information Hypothesis: The Expected Information Hypothesis, 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 $[H(p) - \epsilon, H(p) + \epsilon]$ for a natural language. Text that falls outside of this region is likely perceived as unnatural.

Though both hypotheses state that a "Goldilocks" zone exists around an anchor entropy value(s) for natural-sounding language generation, they differ on key operational details. First, the anchor entropy value in the case of the expected information hypothesis is computed under ancestral sampling. The choice of ancestral sampling (temperature sampling with $temp = 1$) for defining the anchoring entropy value is questionable given that the decoding approach itself performs relatively poorly on text quality (low Mauve score) and contextuality (low F1 score) metrics.

Additionally, the stable entropy hypothesis has an inherent temporal element to it which is missing in the expected information hypothesis. This can be useful to analyze evolving behavior of various decoding algorithms over the course of the generation such as a correlation between repetitions exhibited by the greedy and beam search and the catastrophic drop in entropy, or the ancestral sampling with high temperature (temp=1.5) and steadily increasing entropy.

Stable Entropy Hypothesis and Local Typicality: Meister et al. (2023) define the concept of local typicality and hypothesize that natural-sounding language belongs to the local typical set for the human language process. The authors further use this concept to propose typical sampling, one of the decoding algorithms evaluated in the paper.

Meister et al. (2023) define (T, ϵ)-**locally typical set** of language process $\mathbf{Y}$ ($\mathbf{Y} = \{Y_t\}_{t=1}^{\infty}$) under distribution p, as a set of all sequences of length T such that

$$\mathcal{L}_{\epsilon}^{(T)} = \{ \mathbf{y} = y_0 \cdots y_T | \forall 1 \le t \le T,$$
$$\left| \log p(y_t | \mathbf{y}_{<t}) + H(Y_t | \mathbf{Y}_{<t} = \mathbf{y}_{<t}) \right| < \epsilon \} \quad (4)$$

In words, this means that we should expect every word in natural-sounding sentences to be close to the expected information content under p, i.e., the conditional entropy given prior context. (Meister et al., 2023).

This notion of local typicality differs from the stable entropy hypothesis in two crucial ways. First, local typicality bounds the surprisal or the information content measured using its negative log probability. The stable entropy hypothesis, in contrast, bounds the entropy of the conditional distribution of the model [7]. Second, the stable entropy zone is anchored around the stable entropy baseline which is defined in terms of the entropy of the model under target distribution whereas local typicality uses the 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. This conclusion does not hold as generated text almost always degenerates under greedy decoding in open-ended generation tasks.

5. Conclusion

In this paper, we presented the stable entropy hypothesis which states that the entropy of natural language stays in a narrow zone around the stable baseline which is defined as the mean entropy of the model under the target 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. Next, we leveraged this analysis to propose a mostly-deterministic entropy-aware decoding method. Our dialog and text completion experiments show that entropy-aware decoding is competitive with other decoding methods on quality, and repetitiveness while being more contextual. We hypothesize that the mostly deterministic nature of entropy-aware decoding 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.

6. Acknowlegment

---

[7]Meister et al. (2023) refer to this as conditional entropy.
[8]https://www.calculquebec.ca
[9]https://www.computecanada.ca

References

Basu, S., Ramachandran, G. S., Keskar, N. S., and Varshney, L. R. Mirostat: A Neural Text Decoding Algorithm that Directly Controls Perplexity, January 2021.

Bojar, O. r., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4717.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, June 2020.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2097. URL https://aclanthology.org/N18-2097.

Fan, A., Lewis, M., and Dauphin, Y. 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.

Hamborg, F., Meuschke, N., Breitinger, C., and Gipp, B. 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.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching Machines to Read and Comprehend. *arXiv:1506.03340 [cs]*, June 2015.

Hewitt, J., Manning, C. D., and Liang, P. Truncation Sampling as Language Model Desmoothing, October 2022.

Holtermann, C., Lauscher, A., and Ponzetto, S. 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.

Holtzman, A., Buys, J., Forbes, M., and Choi, Y. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751 [cs]*, April 2019.

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

Krishna, K., Chang, Y., Wieting, J., and Iyyer, M. RankGen: Improving Text Generation with Large Ranking Models, May 2022.

Lee, N., Ping, W., Xu, P., Patwary, M., Shoeybi, M., and Catanzaro, B. Factuality Enhanced Language Models for Open-Ended Text Generation, June 2022.

Levy, R. *Probabilistic models of word order and syntactic discontinuity*. stanford university, 2005.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. 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.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation, January 2020.

Meister, C., Cotterell, R., and Vieira, T. 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.

Meister, C., Wiher, G., Pimentel, T., and Cotterell, R. 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.

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

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 314–319, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5333. URL https://aclanthology.org/W19-5333.

Pang, B., Nijkamp, E., Kryściński, W., Savarese, S., Zhou, Y., and Xiong, C. Long document summarization with top-down and bottom-up inference, 2022.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers, November 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. Compressive Transformers for Long-Range Sequence Modelling. November 2019. doi: 10.48550/arXiv.1911.05507.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., and Weston, J. Recipes for building an open-domain chatbot. *arXiv:2004.13637 [cs]*, April 2020.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., Behrooz, M., Ngan, W., Poff, S., Goyal, N., Szlam, A., Boureau, Y.-L., Kambadur, M., and Weston, J. BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage, August 2022.

Smith, E. M., Williamson, M., Shuster, K., Weston, J., and Boureau, Y.-L. 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.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3450–3466, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.304. URL https://aclanthology.org/2021.findings-acl.304.

Tiedemann, J. and Thottingal, S. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.61.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. 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.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. OPT: Open Pre-trained Transformer Language Models, June 2022.

Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., and Liu, P. J. Calibrating sequence likelihood improves conditional language generation, 2022.

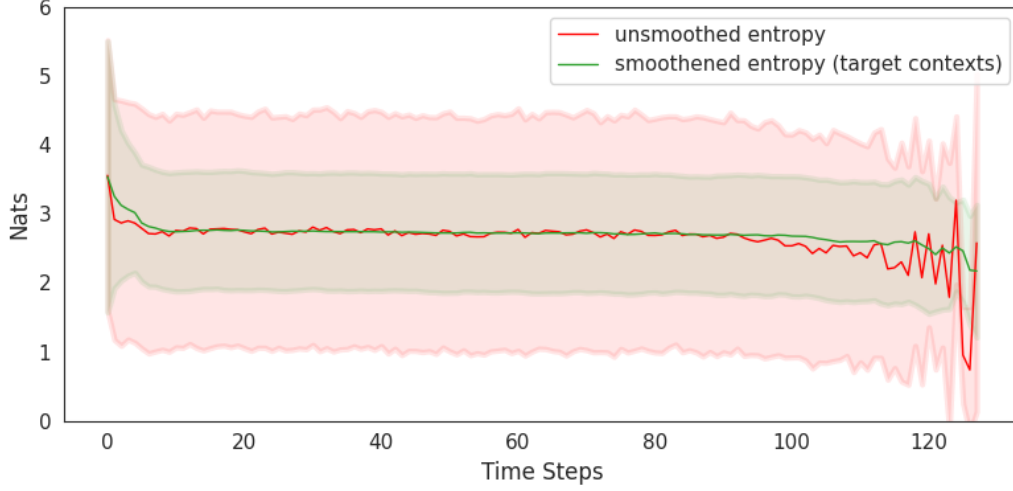## A. Unsmoothed And Smoothed Stable Entropy Baseline and Stable Entropy Zone

Figure 5. **Smoothed vs unsmoothed Stable Entropy Zone.** We plot the stable entropy baseline and the stable entropy zone using both the smoothed and unsmoothed entropy. In either case, we observe that the stable entropy baseline and the stable entropy zone remain mostly flat except for the first and last few time steps.

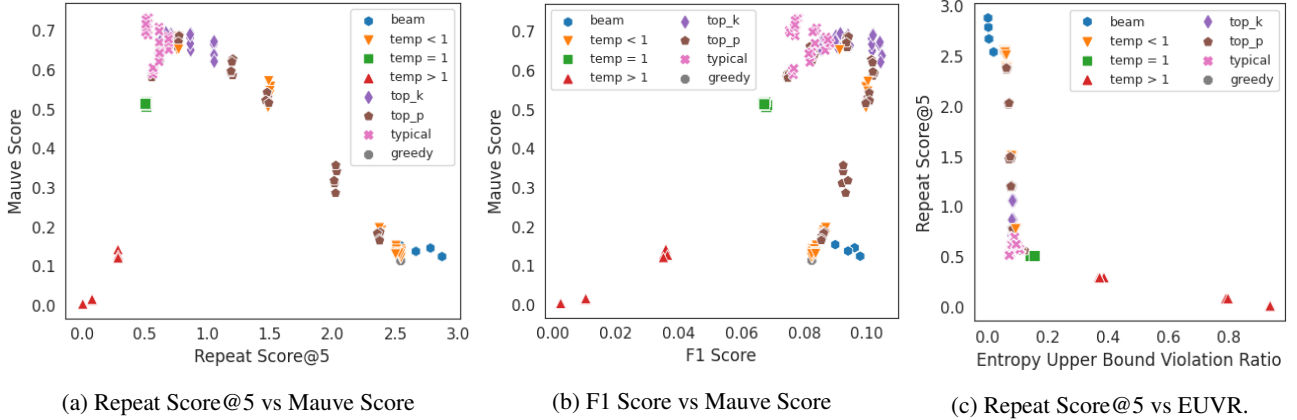## B. Additional Correlational Plots And Configurations for Stable Entropy Hypothesis Analysis

(a) Repeat Score@5 vs Mauve Score

(b) F1 Score vs Mauve Score

(c) Repeat Score@5 vs EUVR.

Figure 6. **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.

C. Quantitative Results for Analysis of Various Decoding Algorithms for Text Completion Setup.

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
Target completions	1.000	0.605	1.381	1.000	0.136	0.0631	0.0731

Table 5. **Quantitiative results for text completion analysis.** F1 score between the target and 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.

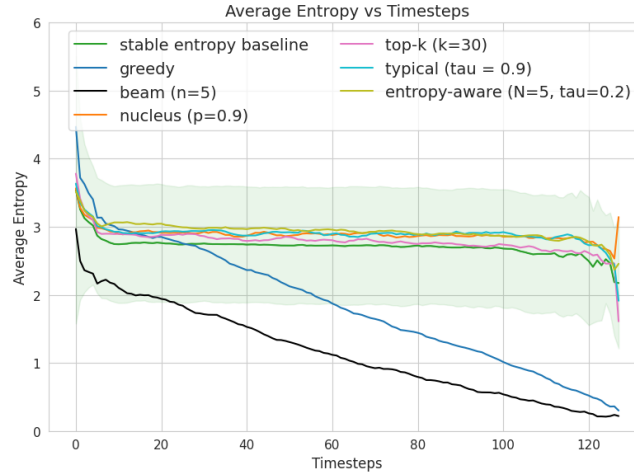# D. Mean Conditional Entropies of Various Decoding Algorithms

Figure 7. **Mean conditional entropies of various decoding algorithms.** We observe that the mean entropy under greedy and beam search drops near-monotonically over the sequence length. Well-tuned Sampling-based methods and entropy-aware decoding nearly follow the stable entropy baseline.

E. Qualitative Examples

E.1. Wikipedia Text Completion

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 6. 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.2. Dialog Generation

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 7. Dialog qualitative examples where beam search produces at least two 3-gram repeats.

E.3. Beam Search With Various Models in Table 2

F. Models and Hyperparameters

F.1. Various Configurations of Decoding Algorithm Evaluated in Section 2.2.1

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.

State of the Art Models in Table 2

- CNN-DM: Zhao et al. (2022)
- ArXiv: Pang et al. (2022)
- WMT: Ng et al. (2019)

G. Repeat Score@5

We modify the Repeat Score@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 Score@5 as

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

where # 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.
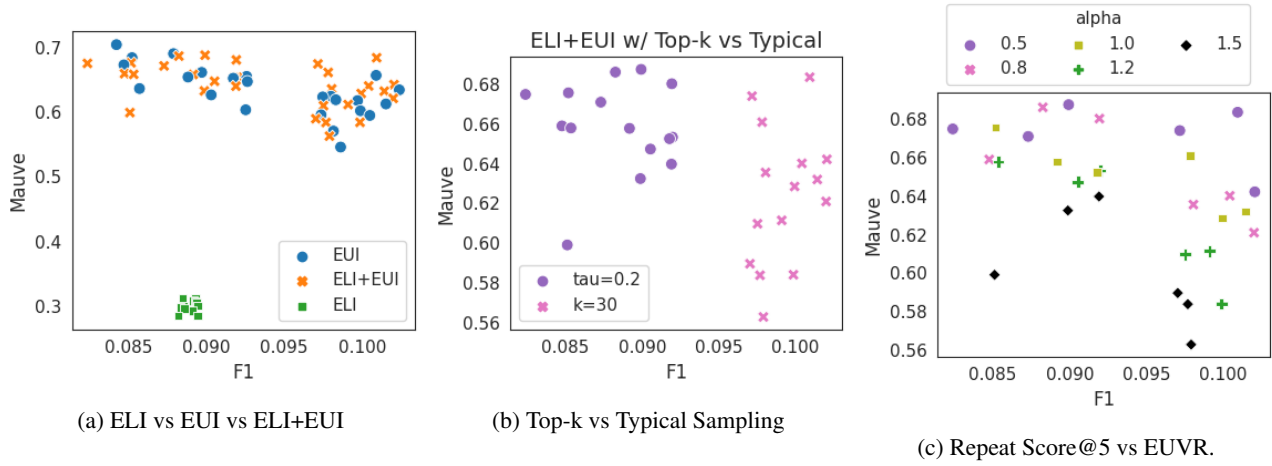
H. Entropy Aware Decoding Analysis

(a) ELI vs EUI vs ELI+EUI (b) Top-k vs Typical Sampling

(c) Repeat Score@5 vs EUVR.

Figure 8. **Entropy-Aware Decoding Analysis.** Figure (a) shows that only entropy lower-bound interventions (ELI) do not work and most gains come from entropy upper-bound interventions. Figure (b) shows that top-k sampling with EAD results in a higher F1 score. Finally, Figure (c) shows that high $alpha$ values result in poorer performance, hence it is reasonable to use the band of ≤ 1.

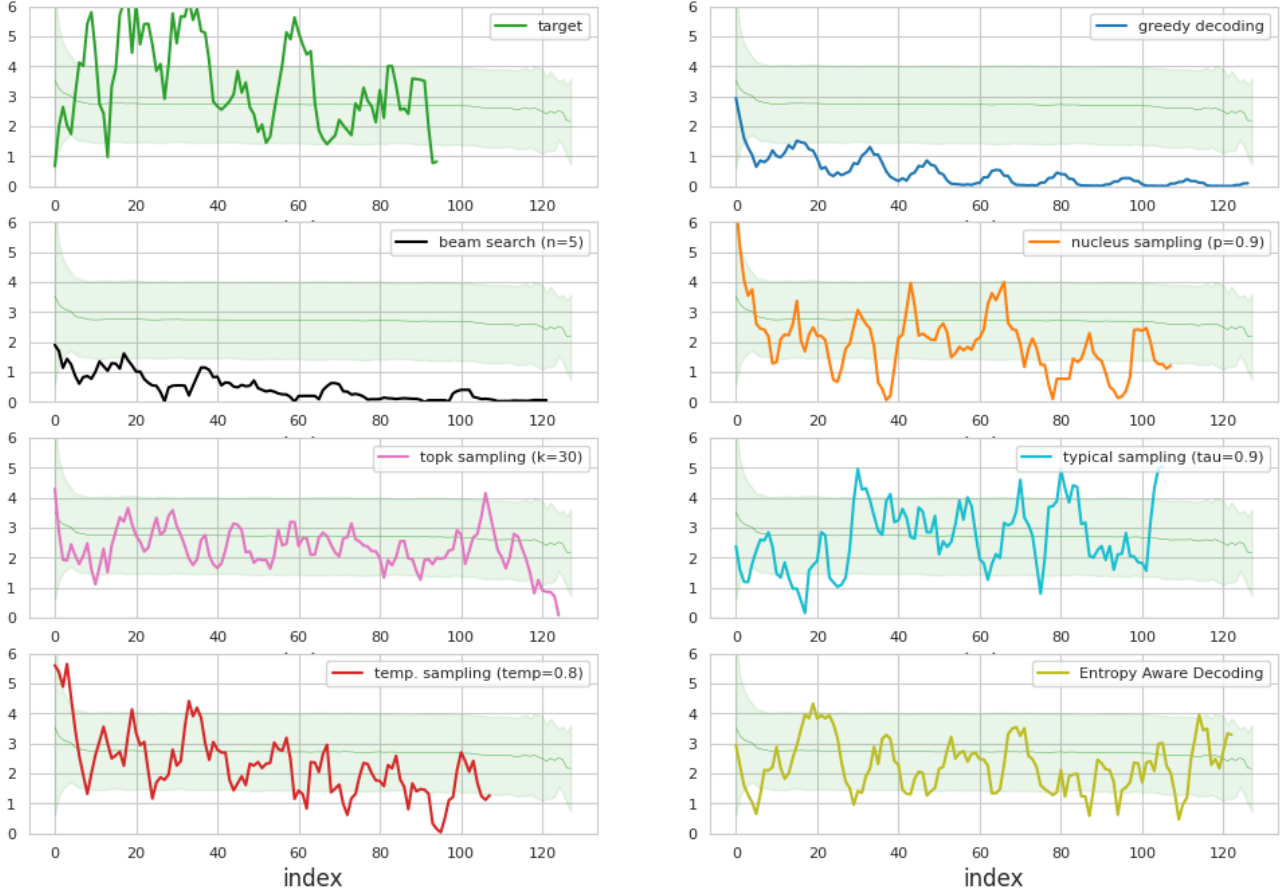# I. Surprisal Under Different Decoding Algorithms

Figure 9. **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 6. The faint green line in the background is the 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.

In the figure, we also find that stochastic decoding methods and entropy-aware decoding induce a context distribution that results in surprisal centered around the target information rate whereas the surprisal under greedy and beam search decoding suffer from a similar catastrophic drop as observed in Figure 3.