

A Systematic Characterization of Sampling Algorithms for Open-ended Language Generation

Moin Nadeem*

Massachusetts Institute of Technology
mnadeem@mit.edu

Kyunghyun Cho

New York University
kyunghyun.cho@nyu.edu

Tianxing He*

Massachusetts Institute of Technology
cloudygoose@csail.mit.edu

James Glass

Massachusetts Institute of Technology
glass@mit.edu

Abstract

This work studies the widely adopted ancestral sampling algorithms for auto-regressive language models, which is not widely studied in the literature. We use the quality-diversity (Q-D) trade-off to investigate three popular sampling algorithms (top- k , nucleus and tempered sampling). We focus on the task of open-ended language generation. We first show that the existing sampling algorithms have similar performance. After carefully inspecting the transformations defined by different sampling algorithms, we identify three key properties that are shared among them: *entropy reduction*, *order preservation*, and *slope preservation*. To validate the importance of the identified properties, we design two sets of new sampling algorithms: one set in which each algorithm satisfies all three properties, and one set in which each algorithm violates at least one of the properties. We compare their performance with existing sampling algorithms, and find that violating the identified properties could lead to drastic performance degradation, as measured by the Q-D trade-off. On the other hand, we find that the set of sampling algorithms that satisfies these properties performs on par with the existing sampling algorithms.¹

1 Introduction

A language model (LM) is a central module for natural language generation (NLG) tasks (Young et al., 2018) such as machine translation (Wu et al., 2018), dialogue response generation (Li et al., 2017), image captioning (Lin et al.), and related tasks. Given a trained LM, finding the best way to generate a sample from it has been an important challenge for NLG applications.

*Equal contribution.

¹Our data and code are available at <https://github.com/moinnadeem/characterizing-sampling-algorithms>.

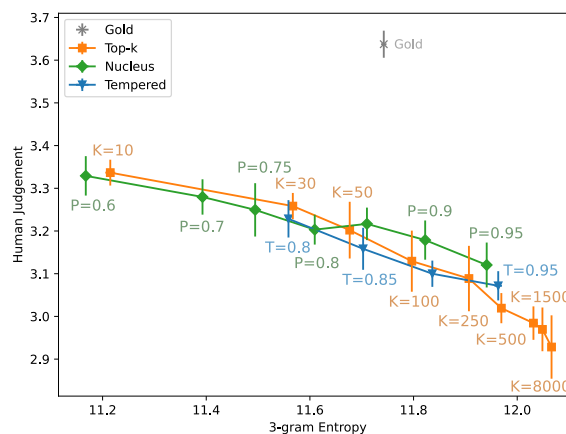


Figure 1: Human evaluation (y-axis: quality, x-axis: diversity, both are the bigger the better) shows that the generation performance of existing sampling algorithms are on par with each other.

Decoding, i.e., finding the most probable output sequence from a trained model, is a natural principle for generation. The beam-search decoding algorithm approximately finds the most likely sequence by performing breadth-first search over a restricted search space. It has achieved success in machine translation, summarization, image captioning, and other subfields.

However, in the task of open-ended language generation (which is the focus of this work), a significant degree of *diversity* is required. For example, conditioned on the prompt “The news says that . . .”, the LM is expected to be able to generate a wide range of interesting continuations. While the deterministic behavior of decoding algorithms could give high-quality samples, they suffer from a serious lack of diversity.

This need for diversity gives rise to a wide adoption of various sampling algorithms. Notably, top- k sampling (Fan et al., 2018), nucleus sampling (Holtzman et al., 2020), and tempered sampling (Caccia et al., 2020) have been used in open-ended

generation (Radford et al., 2018; Caccia et al., 2020), story generation (Fan et al., 2018), and dialogue response generation (Zhang et al., 2020b). However, the sampling algorithm and the hyperparameter are usually chosen via heuristics, and a comprehensive comparison between existing sampling algorithms is lacking in the literature. More importantly, **the underlying reasons behind the success of the existing sampling algorithms still remains poorly understood.**

In this work, we begin by using the quality-diversity (Q-D) trade-off (Caccia et al., 2020) to compare the three existing sampling algorithms. For automatic metrics, we use the BLEU score for quality and n-gram entropy for diversity. We also correlate these automatic metrics with human judgements. The first observation we draw is that top- k , nucleus and tempered sampling perform on par in the Q-D trade-off, as shown in Figure 1. Motivated by this result, we extract three key properties by inspecting the transformations defined by the sampling algorithms: (1) *entropy reduction*, (2) *order preservation* and (3) *slope preservation*. We prove all three properties hold for the three existing sampling algorithms.

We then set out to systematically validate the importance of the identified properties. To do so, we design two sets of new sampling algorithms in which each algorithm either violates one of the identified properties, or satisfies all properties. Using the Q-D trade-off, we compare their efficacy against existing algorithms, and find that violating these identified properties could result in significant performance degradation. More interestingly, we find that the set of sampling algorithms that satisfies these properties has generation performance that matches the performance of existing sampling algorithms.

2 Sampling Algorithms for Autoregressive Language Models

2.1 Autoregressive Language Modeling

The task of autoregressive language modeling is to learn the probability distribution of the $(l + 1)$ -th word W_{l+1} in a sentence W conditioned on the word history $W_{1:l} := (W_1, \dots, W_l)$ and context C . Here, we use $W_i \in V$ to denote a discrete random variable distributed across a fixed vocabulary V . In this work, the vocabulary is constructed on sub-word level (Sennrich et al., 2016).

Given a training set D , maximum likelihood es-

timization (MLE) has been the most popular framework to train an autoregressive LM (Mikolov et al., 2010). MLE training minimizes the negative log-likelihood (NLL) objective below:

$$L_{\text{MLE}} = \frac{1}{|D|} \sum_{(W,C) \in D} -\sum_{l=0}^{L-1} \log P_{\theta}(W_{l+1}|W_{1:l}, C), \quad (1)$$

where θ denotes model parameters, and $P_{\theta}(\cdot | W_{1:l})$ denotes the conditional model distribution of W_{l+1} given a prefix $W_{1:l}$. For simplicity, we assume all sentences are of length L in the formulations. Since this work focuses on sampling from a given model instead of training it, in the rest of the paper, we abbreviate $P_{\theta}(\cdot)$ as $P(\cdot)$ for brevity.

2.2 Existing Sampling Algorithms

Given a trained LM and a context C , an ancestral sampling algorithm seeks to generate a sequence from $P(W|C)$ by sampling token-by-token from a transformed version of $P(W_{l+1}|W_{1:l}, C)$. We now review and formulate three popular sampling algorithms: top- k (Fan et al., 2018), nucleus (Holtzman et al., 2020), and tempered (Ackley et al., 1985; Caccia et al., 2020) sampling.

We view these algorithms as different transformations applied to the distribution $P(W_{l+1}|W_{1:l}, C)$. First, we treat the conditional distribution $P(W_{l+1}|W_{1:l}, C)$ as a *sorted* vector \mathbf{p} of length $|V|$. By sorting, we rearrange the elements such that if $i < j \rightarrow p_i \geq p_j$.² We list the transformations and their intuition below:

Definition 2.1. (Top- k) In top- k sampling, we only sample from the top K tokens:

$$\hat{p}_i = \frac{p_i \cdot \mathbb{1}\{i \leq K\}}{\sum_{j=1}^K p_j}, \quad (2)$$

where $\mathbb{1}$ is the indicator function, and K ($1 \leq K \leq |V|$) is the hyperparameter.

Definition 2.2. (Nucleus) With a hyperparameter P ($0 < P \leq 1$), in nucleus sampling, we sample from the top- P mass of \mathbf{p} :

$$\hat{p}_i = \frac{p'_i}{\sum_{j=1}^{|V|} p'_j}, \quad (3)$$

where $p'_i = p_i \cdot \mathbb{1}\{\sum_{j=1}^{i-1} p_j < P\}$.

²The token indexes are also permuted accordingly.

Definition 2.3. (Tempered) In tempered sampling, the log probabilities are scaled by $\frac{1}{T}$:

$$\hat{p}_i = \frac{\exp(\log(p_i)/T)}{\sum_{j=1}^{|V|} \exp(\log(p_j)/T)}. \quad (4)$$

In this work, we assume $0 < T < 1$, i.e., the distribution is only made sharper³.

We additionally experiment with a combined version of top- k and tempered sampling:

Definition 2.4. (Tempered Top- k) We combine the transformation defined by top- k and tempered sampling:

$$\hat{p}_i = \frac{p'_i}{\sum_{j=1}^{|V|} p'_j}, \quad (5)$$

where $p'_i = \exp(\log(p_i)/T) \cdot \mathbb{1}\{i \leq K\}$. We set $1 \leq K \leq |V|$ and $0 < T < 1$.

Throughout this work we use $\hat{\mathbf{p}}$ to denote the normalized version of the transformed distribution. All algorithms have hyperparameters to control the entropy of the transformed distribution. For example, K in top- k sampling controls the size of the support of the resulting distribution. We will formalize this statement in Property 1 below.

3 Properties of Sampling Algorithms

As we will show in Section 5.1 (also Figure 1), top- k , nucleus and tempered sampling perform on par with each other under our evaluation. This key observation makes us question: *What are the core principles underlying the different algorithms that lead to their similar performance?*

To answer this question, in this section, we identify three core properties that are provably shared by the existing sampling algorithms. We then design experiments to validate their importance.

3.1 Identifying Core Properties

By inspecting the transformations listed in Definition 2.1, 2.2 and 2.3, we extract the following three properties:

Property 1. (Entropy Reduction): The transformation strictly decrease the entropy of the distribution. Formally, $\mathcal{H}(\hat{\mathbf{p}}) < \mathcal{H}(\mathbf{p})$, where $\mathcal{H}(\mathbf{p}) = -\sum_{i=1}^{|V|} p_i \log p_i$.

³One could also use $T > 1$, but it does not work well in practice.

Property 2. (Order Preservation): The order of the elements in the distribution is preserved. Formally, $p_i \geq p_j \rightarrow \hat{p}_i \geq \hat{p}_j$.

Property 3. (Slope Preservation): The ‘‘slope’’ of the distribution is preserved. Formally, $\forall \hat{p}_i > \hat{p}_j > \hat{p}_k > 0$ (i.e., they are not truncated), we have $\frac{\log p_i - \log p_j}{\log p_j - \log p_k} = \frac{\log \hat{p}_i - \log \hat{p}_j}{\log \hat{p}_j - \log \hat{p}_k}$.

The order preservation property implies that truncation can only happen in the tail of the distribution, which aligns with top- k and nucleus sampling. The slope preservation property is stronger than the order preservation property in that not only the ordering, but also the relative magnitude of the elements in the distribution needs to be somewhat preserved by the transformation.

All these three properties are shared by the three existing sampling algorithms:

Proposition 1. Property 1, 2 and 3 hold for the top- k , nucleus and tempered sampling transformations formulated in Definitions 2.1, 2.2 and 2.3.

Proof. See Appendix B. \square

We then set out to validate the importance of these identified properties in the aspects of *necessity* and *sufficiency*. To do so, we design two sets of new sampling algorithms in which each algorithm either violates one of the identified properties, or satisfies all properties. We list them in the next section.

3.2 Designed Sampling Algorithms

Property-violating algorithms To validate the necessity of each property, we design several sampling algorithms which *violate at least one of the identified properties*. In our experiments, we check whether that violation leads to a significant degradation in performance. We list them below:

Definition 3.1. (Target Entropy) Based on tempered sampling, target entropy sampling tunes the temperature t such that the transformed distribution has entropy value equal to the hyperparameter E ($0 < E \leq \log |V|$). We formulate it below:

$$\hat{p}_i = \frac{\exp(\log(p_i)/t)}{\sum_{j=1}^{|V|} \exp(\log(p_j)/t)}, \quad (6)$$

where t is selected such that $H(\hat{\mathbf{p}}) = E$.

Target entropy sampling violates entropy reduction, because when $H(\mathbf{p}) < E$, the entropy will be tuned up (i.e., $H(\hat{\mathbf{p}}) > H(\mathbf{p})$).

Definition 3.2. (Random Mask) In random mask sampling, we randomly mask out tokens in the distribution with rate R . We formulate it below:

$$\hat{p}_i = \frac{p'_i}{\sum_{j=1}^{|V|} p'_j}, \quad (7)$$

where $p'_i = p_i \cdot \mathbb{1}\{i = 1 \text{ or } u_i > R\}$ and $u_i \sim U(0, 1)$. The hyperparameter R ($0 < R \leq 1$) controls the size of the support of the resulting distribution. In Appendix A, we show it is crucial that the token which is assigned the largest probability (p_1) is never be masked.

Random mask sampling is different from top- k or nucleus sampling in that the masking not only happens in the tail of the distribution. Therefore, it violates the order preservation property.

Definition 3.3. (Noised Top- k) We add a *sorted* noise distribution to the result from top- K transformation, and the weight of the noise distribution is controlled by a hyperparameter W ($0 \leq W \leq 1$). We formulate it below:

$$\hat{\mathbf{p}} = (1 - W)\hat{\mathbf{p}}^{\text{top-K}} + W\mathbf{p}^{\text{noise-K}}, \quad (8)$$

where $\mathbf{p}^{\text{noise-K}}$ is a uniformly sampled *sorted* K -simplex, which satisfies $\sum_{i=1}^K p_i^{\text{noise-K}} = 1$ and $i < j \rightarrow p_i^{\text{noise-K}} \geq p_j^{\text{noise-K}} \geq 0$.

The sorted nature of the noise distribution $\mathbf{p}^{\text{noise-K}}$ maintains order preservation. However, it violates slope preservation, and the noise weight W controls the degree of the violation.

Property-satisfying algorithms To validate the sufficiency of the identified properties, we design two new sampling algorithms for which *all three properties hold*. And in our experiments we check whether their performance is on par with the existing sampling algorithms. We list them below:

Definition 3.4. (Random Top- k) We design a randomized version of top- k sampling: At each time step, we sample a uniformly random float number $u \sim U(0, 1)$, and use it to specify a top- k truncation:

$$\hat{p}_i = \frac{p_i \cdot \mathbb{1}\{i \leq k\}}{\sum_{j=1}^k p_j}, \quad (9)$$

where $k = \lfloor 1 + M \cdot u \rfloor$. The hyperparameter M ($1 \leq M < |V|$) controls the maximum truncation threshold.

Definition 3.5. (Max Entropy) Max entropy sampling is similar to target entropy sampling (Definition 3.1). However to match entropy reduction (Property 1), we only tune the temperature when $\mathcal{H}(\mathbf{p}) > E$, where E is the hyperparameter ($0 < E \leq \log |V|$):

$$\hat{p}_i = \begin{cases} \frac{\exp(\log(p_i)/t)}{\sum_{j=1}^{|V|} \exp(\log(p_j)/t)}, & \text{if } \mathcal{H}(\mathbf{p}) > E \\ p_i, & \text{otherwise} \end{cases}, \quad (10)$$

where t is selected so that $\mathcal{H}(\hat{\mathbf{p}}) = E$.

It is easy to prove that Property 1, 2, and 3 holds for the transformations defined by random top- k and max entropy sampling, and we omit the proof for brevity.

4 Experiment Setup

In this section, we first establish evaluation protocols, and then describe the model and data we use for the open-ended language generation task.

4.1 Evaluation via the Q-D Trade-off

How to efficiently measure the generation performance of a NLG model has been an important open question. Most existing metrics either measure the *quality* aspect (e.g. BLEU score) or the *diversity* (e.g. n-gram entropy) aspect. To make the situation more complicated, each sampling algorithm has its own hyperparameters which controls the trade-off between quality and diversity.

To address the challenges above, we adopt the quality-diversity trade-off proposed by Caccia et al. (2020). In the Q-D trade-off, we perform a fine-grained sweep of hyperparameters for each sampling algorithm, and compute the quality and diversity score for each configuration. We report two pairs of Q/D metrics, with one pair using automatic evaluation and the other using human evaluation. In the next two sections, we describe the metrics we use, and refer readers to Caccia et al. (2020) for more intuition behind the Q-D trade-off.

4.1.1 Automatic Evaluation

For automatic metrics, we adopt the corpus-BLEU (Yu et al., 2016) metric to measure quality and the self-BLEU (Zhu et al., 2018) metric to measure diversity. We formulate them below.

Given a batch of generated sentences S_{gen} and a batch of sentences from ground-truth data as references S_{ref} , corpus-BLEU returns the average

BLEU score (Papineni et al., 2002) of every model generated sentence against the reference set:

$$\text{corpus-BLEU}(S_{\text{gen}}, S_{\text{ref}}) = \frac{1}{|S_{\text{gen}}|} \sum_{W \in S_{\text{gen}}} \text{BLEU}(W, S_{\text{ref}}). \quad (11)$$

A higher corpus-BLEU score means that the generated sequences has better quality in that it has higher ngram-level overlap with the reference data. Based on the same intuition, we define the self-BLEU metric to quantify the diversity aspect:

$$\text{self-BLEU}(S_{\text{gen}}) = \text{corpus-BLEU}(S_{\text{gen}}, S_{\text{gen}}), \quad (12)$$

where a lower self-BLEU score means that the samples have better diversity.

In our experiments, we feed the first ten subwords of every sample from test set to the model, and compare the model-generated sequences to the reference samples in the validation set. We use 10,000 samples to compute corpus-BLEU or self-BLEU, i.e., $|S_{\text{gen}}| = |S_{\text{ref}}| = 10,000$.

Automatic evaluation enables us to do a fine-grained sweep of the hyperparameters for each sampling algorithm, and compare them in the quality-diversity trade-off. However, observations from automatic evaluation could be misaligned with human evaluation (Belz and Reiter, 2006). Therefore, we confirm our key observations with human evaluation.

4.1.2 Human Evaluation

Quality We ask a pool of 602 crowdworkers on Amazon Mechanical Turk to evaluate various sampling configurations in the quality aspect. Each worker is presented a set of ten samples along with the prompts (prefixes). They are then asked to rate how likely the sentence would appear in a news article between 0 and 5 (Invalid, Confusing, Unspecific, Average, Expected, and Very Expected respectively).

We focus on the Gigaword dataset for human evaluation since news articles are ubiquitous and do not often require expert knowledge for quality judgement. For each configuration (sampling algorithm and hyperparameter pair) we ask crowdworkers to rate 200 samples in total. To get an accurate rating for each sample, we enlist 25 different crowdworkers to rate each sample. We report mean and standard deviation from 5 independent runs (each with 40 samples) as error bar.

By manual inspection, we find that the time spent in the annotations is a good indicator of the quality

of the rating. Therefore, we estimate the human judgement score for a sample as the average rating of the 20 crowdworkers (out of 25) who took the most time to rate the samples. We provide further details about our setup in Appendix C and D.

Diversity It is difficult for human annotators to estimate diversity of text (Hashimoto et al., 2019). Therefore, we use the *n-gram entropy* metric (Zhang et al., 2018; He and Glass, 2019). Given S_{gen} which contains a large number of samples, we measure its diversity using the following formulation:

$$\mathcal{H}^{n\text{-gram}}(S_{\text{gen}}) = \sum_{g \in G_n} -r(g) \log r(g), \quad (13)$$

where G_n is the set of all n -grams that appeared in S_{gen} , and $r(g)$ refers to the ratio (frequency) of n -gram g w.r.t. all n -grams in the S_{gen} . For the estimation of n -gram entropy, we generate 50,000 samples from each sampling configuration.

We will report human quality score either paired with n -gram entropy or with self-BLEU as diversity metric. We find they give similar observations.

4.2 Model and Datasets

We separately fine-tune GPT2-small (Radford et al., 2018; Wolf et al., 2019) (110M parameters) on the Gigaword (Graff et al., 2003; Napoles et al., 2012) and the Wikitext-103 (Merity et al., 2017) datasets. We use the same tokenization as GPT-2, and add additional padding and end-of-sequence tokens (`[EOS]`) to the sentences.

To generate a sequence, we feed a length-10 prefix from test data into the fine-tuned GPT-2 model, and use a sampling algorithm to complete the sentence. Since shorter samples are more difficult to judge in quality (Ippolito et al., 2020), we filter all generated sentence completions to be between 40 and 50 subwords, and filter our validation and test set to meet the same requirements. To permit validation and test sets that are large enough to prefix 10,000 sentences for the corpus-BLEU metric, we re-chunk the first 80% of the Gigaword dataset for the training set, 15% for validation, and the last 5% for the test set. Similarly, we re-chunk the first 97% of the Wikitext-103 dataset for training, and leave 1.5% for validation and 1.5% for test.

5 Empirical Results

First, we compare existing sampling algorithms, and then move on to validate the necessity and

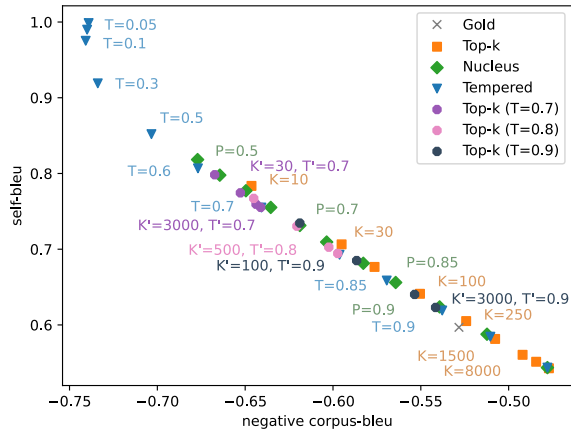


Figure 2: The performance (x-axis: quality, y-axis: diversity, both are the smaller the better) of top- k , nucleus, tempered and tempered top- k sampling are on par on the Gigaword dataset, as shown by automatic evaluation.

sufficiency of the identified properties.

5.1 Comparison of Existing Algorithms

We compare top- k , nucleus, and tempered sampling via automatic and human evaluation. We do a fine-grained sweep of hyperparameters for each sampling algorithm on the Gigaword dataset. The results are shown in Figure 1 (human evaluation) and Figure 2 (automatic evaluation). We also show the quality and diversity score for human text in the test data for reference, which is labeled as gold.

Both automatic and human evaluations demonstrate that the performance of top- k , nucleus and tempered sampling are on par with each other, with no significant gap. When the hyperparameters (K , P and T) are tuned so that different sampling has the same diversity (measured by self-BLEU or n-gram entropy), their quality (measured by corpus-BLEU or human rating) are close.

Additionally, we compare tempered top- k sampling with the existing algorithm also in Figure 2. We find that adding the tempered transformation only moves top- k sampling along the Q-D trade-off, instead of yielding a better or a worse sampling algorithm. For example, the performance of the $K = 500, T = 0.8$ configuration for tempered top- k sampling is very close to the $K = 30$ configuration for the top- k sampling.

Motivated by these observations, we identify three core properties (elaborated in Section 3.1) that are shared among the sampling algorithms: *entropy reduction*, *order preservation* and *slope preservation*. In the following two sections, we

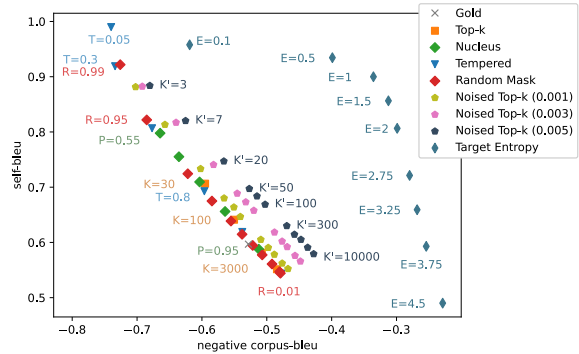


Figure 3: Automatic evaluation of the noised top- k , target entropy, and random mask sampling proposed to validate the necessity of the identified properties. The results show that violation of entropy reduction and slope preservation could lead to drastic performance degradation, while the order preservation property could be further relaxed.

present experiments validating the necessity or sufficiency aspect of the properties.

5.2 Property-violating Algorithms

In Figure 3, we compare the generation performance of the property-violating sampling algorithms (designed in Section 3.2), against the existing algorithms using automatic evaluation on the Gigaword dataset. We make the following observations: First, the target entropy sampling, which violates entropy reduction, has significantly worse performance; Second, even with small noise weight W , the performance of noised top- k sampling degrades from the original top- k sampling, and the gap becomes larger as W increases; Last, the random mask sampling is on par with the existing sampling algorithms in performance. We further confirm this observation with human evaluation in Figure 5.

These results suggest that the violation of entropy reduction or slope preservation could lead to drastic performance degradation. On the other hand, the competitive performance of random mask sampling suggests that order preservation could be further relaxed.

In the next section, we investigate the sufficiency aspect of the identified properties.

5.3 Property-satisfying Algorithms

We now compare the generation performance of the property-satisfying sampling algorithms (designed in Section 3.2) with the existing sampling algorithms. The results from the Gigaword dataset

Sampling	Conditional Samples
Existing Sampling Algorithms	
<i>Top-k</i> ($K = 30$)	<i>steven spielberg's dreamworks movie studio</i> said monday it was filing a lawsuit, accusing us studio executives of defrauding hundreds of thousands of dollars in refunds and other damages.
<i>Nucleus</i> ($P = 0.80$)	<i>steven spielberg's dreamworks movie studio</i> has failed to attract the kind of business and development investors that jeffrey hutchinson dreamed up in the past.
<i>Tempered</i> ($T = 0.85$)	<i>steven spielberg's dreamworks movie studio</i> plans to spend the rest of the year producing the high-speed thriller "the earth's path" and an upcoming sequel, the studio announced on wednesday.
Property-satisfying Sampling Algorithms	
<i>Random Top-k</i> ($R = 90$)	<i>steven spielberg's dreamworks movie studio</i> is planning to make a movie about a young man who is a <unk>, a man who has a dream of being the first man to be born with the ability to walk on water.
<i>Max Entropy</i> ($E = 2.75$)	<i>steven spielberg's dreamworks movie studio</i> has agreed to pay \$ ##.# million to director john nichols (£ ##.# million, ###, a record in the studio circulation), the studio announced sunday..
Property-violating Sampling Algorithms	
<i>Random Mask</i> ($R = 0.75$)	<i>steven spielberg's dreamworks movie studio</i> scored a big win with a \$ ##.## million (euro ##.## million) direct-to-video (dvds) deal to develop the ##### short story "the rose garden".
<i>Noised Top-k</i> ($K=50, W=5e-3$)	<i>steven spielberg's dreamworks movie studio</i> is in disarray and has a few directors and a lot of stock involved, leaving it only a matter of time before spielberg's departure from the nobel peace prize .
<i>Target Entropy</i> ($E = 2.75$)	<i>steven spielberg's dreamworks movie studio</i> production scored an action boost m boom, nabbing an 'd after the ##th instal specialization with nominations of fritz, ika, ivan english ape and evlyn mcready.

Table 1: Generated sequences with the same prefix *steven spielberg's dreamworks movie studio* by different sampling algorithms. The hyperparameters are chosen such that the algorithms yield roughly the same diversity measured by self-BLEU. The poor-quality spans are highlighted in red.

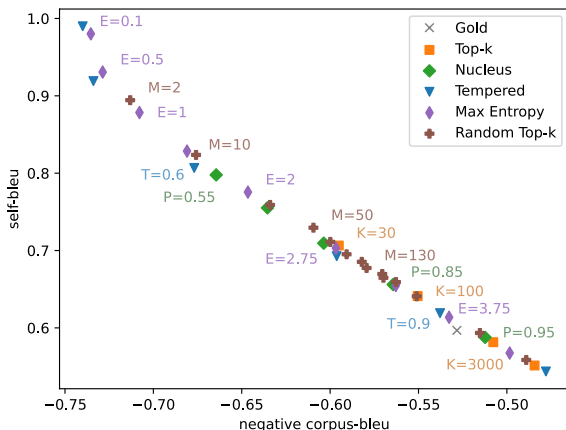


Figure 4: The proposed random top- k and max entropy schedulers, which meet the identified properties, are on par in performance with existing methods in automatic evaluation on the Gigaword dataset.

are shown in Figure 3 (for automatic evaluation) and Figure 5 (for human evaluation). For completeness, we also replicate Figure 5 with self-BLEU as the diversity measure in Appendix F. We also present results from automatic evaluation on the Wikitext-103 dataset in Figure 6, with consistent observations.

The evaluations consistently show that the performance of random top- k and max entropy sampling (and random mask sampling in last section) is on par with top- k , nucleus, and tempered sampling. These results strengthen the importance of the iden-

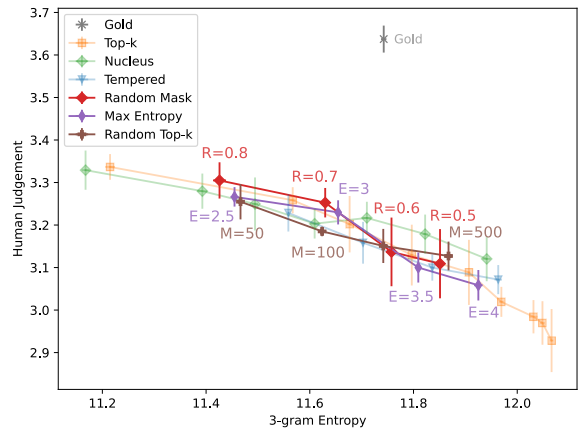


Figure 5: Human evaluation also shows that the proposed sampling algorithms has performance on par with the existing methods on the Gigaword dataset. Appendix F repeats this plot with self-BLEU.

tified properties in that, new sampling algorithms could get competitive generation performance as long as they meet the identified properties.

5.4 Qualitative Analysis

We list samples from the proposed sampling algorithms and compare them with the existing ones in Table 1. We choose the hyperparameter of each sampling algorithm so that each algorithm exhibits a similar level of diversity (as measured by self-BLEU). By manual inspection, we find that the quality of samples from property-satisfying sam-

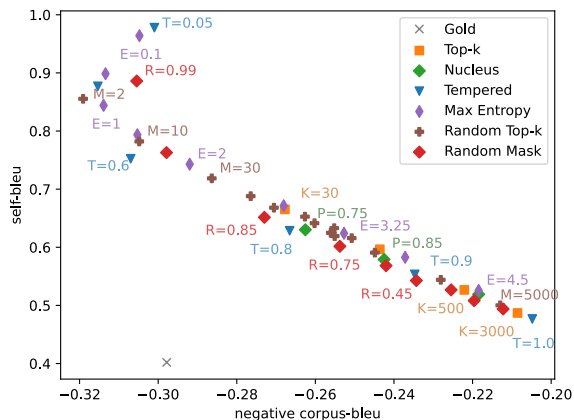


Figure 6: Automatic evaluation on the Wikitext-103 dataset: The performance of proposed sampling algorithms are on par with top- k , nucleus, and tempered sampling.

pling algorithms is on par with samples from the existing algorithms. In particular, the samples from random top- k , max entropy, and random masked sampling are all coherent and informative.

In contrast, the samples from noised top- k and target entropy algorithms, tend to be less semantically and syntactically coherent. In particular, the target entropy sampling algorithm, which obtains the lowest quality score measured by corpus-BLEU, lacks basic language structure. In comparison to target entropy, noised top- k is syntactically coherent, but exhibits logical and factual inconsistencies. These observations aligns with the results we get from automatic evaluation.

6 Related Works

Despite the popularity of sampling algorithms in natural language generation, a rigorous comparison or scrutiny of existing algorithms is lacking in the literature. Ippolito et al. (2019) provides a comparison between sampling and decoding algorithms. Holtzman et al. (2020) proposes nucleus sampling, and compare it with top- k sampling (Fan et al., 2018). However, only a few hyperparameter configurations are tested. In Hashimoto et al. (2019) and Caccia et al. (2020), temperature sampling is used and the hyperparameter T is tuned to trade-off between diversity and quality, but it lacks comparisons with other sampling algorithms. Welleck et al. (2020) studies the *consistency* of existing sampling and decoding algorithms, without comparing the generation performance.

In this work we mainly use the quality-diversity trade-off (Caccia et al., 2020) to conduct a compar-

ison of different sampling algorithms. Parallel to our work, Zhang et al. (2020a) also uses the quality-diversity trade-off to compare top- k , nucleus, and tempered sampling. Their observation is similar to ours: The performance of the existing algorithms are close with no significant gap.

More importantly, the underlying reasons for the success of various sampling algorithms remain poorly understood. Zhang et al. (2020a) proposes the *selective* sampling algorithm, which fails to outperform existing approaches. This failed attempt suggests the need for a better understanding of the strengths and weaknesses of existing methods. To the best of our knowledge, our work provides the first systematic characterization of sampling algorithms, where we attribute the success of existing sampling algorithms to a shared set of properties. We show that we can propose novel sampling algorithms based on the identified properties, and reach competitive generation performance as measured by both automatic and human evaluation.

7 Limitations and Future Work

Our core contribution is the three properties of sampling algorithms that we conjecture are crucial for competitive generation performance. While we design a set of experiments to validate their necessity and sufficiency, the observations we make are still empirical. We emphasize that **it is completely possible that there exists some crucial property, that is yet to be discovered, and can lead to significantly better generation performance**. Therefore, the exploration of novel sampling algorithms (Zhang et al., 2020a) should still be encouraged.

On the other hand, to provide a comprehensive study, we focus on the open-ended language generation task with the GPT-2 model. As future work, it would be interesting to check whether our observations also hold on other tasks such story generation or dialogue response generation, or with weaker language models in low-resource setting.

8 Conclusion

This work studies sampling algorithms for the open-ended language generation task. We show that the existing algorithms, namely top- k , nucleus, and tempered sampling, have similar generation performance as measured by the quality-diversity trade-off evaluation. Motivated by this result, we identify three key properties that we prove are shared by

the existing algorithms. To validate the importance of these identified properties, we design a set of new sampling algorithms, and compare their performance with the existing sampling algorithms. We find that violation of the identified properties may lead to drastic performance degradation. On the other hand, we propose several novel algorithms, namely random top- k and max entropy sampling, that meet the identified properties. We find that their generation performance is on par with the existing algorithms.

Acknowledgments

The authors sincerely thank Yixin Tao, Jingzhao Zhang and Yonatan Belinkov for useful discussions. This work was partly supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI), Samsung Electronics (Improving Deep Learning using Latent Structure). Kyunghyun Cho thanks CIFAR, Naver, eBay, NVIDIA and Google for their support.

This research was sponsored in part by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation herein.

References

- David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. *A Learning Algorithm for Boltzmann Machines*, volume 9, pages 147–169.
- Anja Belz and Ehud Reiter. 2006. [Comparing automatic and human evaluation of NLG systems](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.
- Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. [Language gans falling short](#). In *Proceedings of the International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.
- Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. [Unifying human and statistical evaluation for natural language generation](#). In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tianxing He and James R. Glass. 2019. [Negative training for neural dialogue response generation](#). *CoRR*, abs/1903.02134.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation](#). In *Proceedings of the International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. [Automatic detection of generated text is easiest when humans are fooled](#). In *Proceedings of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.
- Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. [Comparison of diverse decoding methods from conditional language models](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.
- Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. [Adversarial learning for neural dialogue generation](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proceedings of European Conference on Computer Vision*.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. [Pointer sentinel mixture models](#). In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.
- Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. [Recurrent neural network based language model](#). In *Proceedings of the International Speech Communication Association*, pages 1045–1048.

- Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. [Annotated Gigaword](#). In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. [Language models are unsupervised multitask learners](#).
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Sean Welleck, Ilya Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. 2020. [Consistency of a recurrent language model with respect to incomplete decoding](#). *CoRR*, abs/2002.02492.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#). *CoRR*, abs/1910.03771.
- Lijun Wu, Yingce Xia, Fei Tian, Li Zhao, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. [Adversarial neural machine translation](#). In *Proceedings of The Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pages 534–549. PMLR.
- Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. [Recent trends in deep learning based natural language processing \[review article\]](#). *IEEE Comput. Intell. Mag.*, 13(3):55–75.
- Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2016. [Seqgan: Sequence generative adversarial nets with policy gradient](#). *CoRR*, abs/1609.05473.
- Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020a. [Trading off diversity and quality in natural language generation](#).
- Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. [Generating informative and diverse conversational responses via adversarial information maximization](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Proceedings of Neural Information Processing Systems 31*, pages 1810–1820. Curran Associates, Inc.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Texygen: A benchmarking platform for text generation models](#). In *Proceedings of the Conference on Research & Development in Information Retrieval*, pages 1097–1100. ACM.

A Auxiliary Plots

We show the importance of preserving the token with the largest probability (p_1) in the proposed random mask sampling. For comparison, we relax the constraint and define the *random mask-all* sampling:

Definition A.1. (Random Mask-all) The only difference between random mask-all sampling and random mask sampling is that we allow the p_1 token to be masked. We formulate it below:

$$\hat{p}_i = \frac{p'_i}{\sum_{j=1}^{|V|} p'_j}, \quad (14)$$

where $p'_i = p_i \cdot \mathbb{1}\{u_i > R\}$ and $u_i \sim U(0, 1)$.

In Figure 7, we show that if p_1 is allowed to be masked, the generation performance will be seriously degraded.

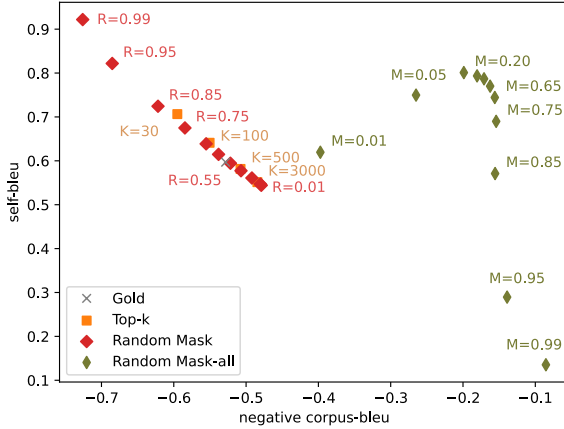


Figure 7: The random mask-all sampling, where p_1 is allowed to be masked, is shown to have worse performance than the random mask sampling. The dataset is Gigaword.

B Proof for Proposition 1

In this section we prove Proposition 1.

Firstly, it is straightforward to prove that Property 2 (order preservation) holds for the top- k , nucleus and tempered sampling and we omit the proof here.

For Property 3 (slope preservation), it holds trivially for nucleus and top- k sampling. We prove it for tempered sampling in the following lemma:

Lemma B.1. Property 3 holds for tempered sampling (Definition 2.3).

Proof. Remember that the tempered sampling with hyperparameter T defines the follow transformation: $\hat{p}_i = \frac{p'_i}{\sum_j p'_j}$, where $p'_i = \exp(\log(p_i)/T)$.

We set $Z = \sum_j p'_j$, then $\forall \hat{p}_i > \hat{p}_j > \hat{p}_k > 0$ we have

$$\begin{aligned} & \frac{\log \hat{p}_i - \log \hat{p}_j}{\log \hat{p}_j - \log \hat{p}_k} \\ &= \frac{\log p'_i - \log p'_j - \log Z + \log Z}{\log p'_j - \log p'_k - \log Z + \log Z} \\ &= \frac{\log p'_i - \log p'_j}{\log p'_j - \log p'_k} \quad (\log Z \text{ is cancelled}) \quad (15) \\ &= \frac{\log(p_i)/T - \log(p_j)/T}{\log(p_j)/T - \log(p_k)/T} \\ &= \frac{\log(p_i) - \log(p_j)}{\log(p_j) - \log(p_k)} \end{aligned}$$

□

Only Property 1 (entropy reduction) is left. We now prove it holds for top- k / nucleus sampling:

Lemma B.2. Property 1 holds for transformations defined by top- k or nucleus sampling (Definition 2.1 and 2.2).

Proof. We first consider the change of entropy when the token with the smallest probability ($p_{|V|}$) is removed from the original distribution ($\hat{p}_i = \frac{p_i}{\sum_{j=1}^{|V|-1} p_i}$, $1 \leq i < |V|$):

$$\begin{aligned} -\mathcal{H}(\mathbf{p}) &= \sum_{i=1}^V p_i \log p_i \\ &= \sum_{i=1}^{V-1} p_i \log p_i + p_{|V|} \log p_{|V|} \\ &= (1 - p_{|V|}) \sum_{i=1}^{V-1} \frac{p_i}{1 - p_{|V|}} \log p_i + p_{|V|} \log p_{|V|} \\ &= \sum_{i=1}^{V-1} \frac{p_i}{1 - p_{|V|}} \log \frac{p_i}{1 - p_{|V|}} + \underbrace{\log(1 - p_{|V|})}_{<0} \\ &\quad + p_{|V|} \left(\log p_{|V|} - \sum_{i=1}^{V-1} \frac{p_i}{1 - p_{|V|}} \log p_i \right) \\ &< \sum_{i=1}^{V-1} \hat{p}_i \log \hat{p}_i + p_{|V|} \left(\log p_{|V|} - \sum_{i=1}^{V-1} \frac{p_i}{1 - p_{|V|}} \log \frac{p_i}{>p_{|V|}} \right) \\ &< \sum_{i=1}^{V-1} \hat{p}_i \log \hat{p}_i + p_{|V|} \left(\log p_{|V|} - \underbrace{\sum_{i=1}^{V-1} \frac{p_i}{1 - p_{|V|}} \log p_{|V|}}_{=\log p_{|V|}} \right) \\ &= \sum_{i=1}^{V-1} \hat{p}_i \log \hat{p}_i = -\mathcal{H}(\hat{\mathbf{p}}) \quad (16) \end{aligned}$$

Therefore, we get $\mathcal{H}(\hat{\mathbf{p}}) < \mathcal{H}(\mathbf{p})$.

By induction (iteratively removing the last token), it is now easy to see that the top- k or nucleus

transformation strictly decrease the entropy of the sampling distribution. \square

Finally, we prove Property 1 (entropy reduction) holds for tempered sampling:

Lemma B.3. Property 1 holds for the transformation defined by tempered sampling (Definition 2.3).

Proof. For convenience, we first rewrite the Temperature transformation:

$$\hat{p}_i = p_i^\alpha = \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \quad (17)$$

where $e_i = -\log(p_i)$ and $\alpha = \frac{1}{T}$. The entropy can be written as:

$$\begin{aligned} \mathcal{H}(\mathbf{p}^\alpha) &= -\sum_i \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \log \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \\ &= \log \sum_j \exp(-\alpha e_j) + \alpha \sum_i e_i \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \end{aligned} \quad (18)$$

Next, we take derivative w.r.t α :

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial \alpha} &= -\underbrace{\sum_i e_i \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)}}_{=0} + \sum_i e_i \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \\ &+ \alpha \frac{\partial}{\partial \alpha} \sum_i e_i \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \\ &= \alpha \sum_i e_i \underbrace{\left[\frac{\partial}{\partial \alpha} \log \frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \right]}_{\text{log-derivative trick}} \left[\frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \right] \\ &= \alpha \sum_i e_i \left[-e_i + \sum_{j'} e_{j'} \frac{\exp(-\alpha e_{j'})}{\sum_j \exp(-\alpha e_j)} \right] \\ &\quad \left[\frac{\exp(-\alpha e_i)}{\sum_j \exp(-\alpha e_j)} \right] \\ &= -\alpha \mathbb{E}_{p^\alpha} \left[e_i^2 - e_i \mathbb{E}_{p^\alpha} [e_i] \right] \\ &= -\underbrace{\alpha}_{>0} \underbrace{\left(\mathbb{E}_{p^\alpha} [e_i^2] - \mathbb{E}_{p^\alpha} [e_i]^2 \right)}_{=\text{Var}_{p^\alpha} [e_i] \geq 0} \\ &< 0 \end{aligned} \quad (19)$$

We can now easily get $\frac{\partial \mathcal{H}}{\partial T} = \frac{\partial \mathcal{H}}{\partial \alpha} \frac{\partial \alpha}{\partial T} > 0$. Therefore, when we apply a tempered transformation with $T < 1$, the entropy will strictly decrease comparing to the original distribution (where $T = 1$). \square

C Mechanical Turk Setup

Our crowdworkers were required to have a HIT acceptance rate higher than 95%, and be located

in the United States. In total, 602 crowdworkers completed our tasks. In order to ensure that we had quality data, we filtered the crowdworker annotations for workers that spent at least 45 seconds on the aggregate task (or 4.5 seconds rating each sentence). 51 crowdworkers were filtered out through this process. Screenshots of our instructions and task are available in Figure(s) 8 and 9 respectively.

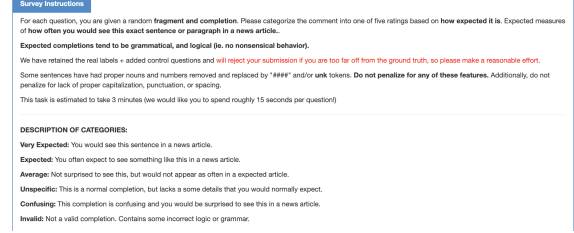


Figure 8: Our instructions for crowdworker task.

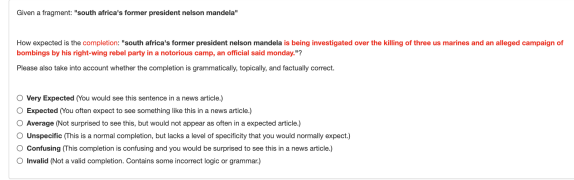


Figure 9: An example of the task given to crowdworkers.

D Convergence of Human Evaluation

When we conduct human evaluation, we provide crowdworkers with 200 generated samples for some configuration, and ask 25 different crowdworkers to evaluate the same sample. However, a reasonable question is whether our human evaluations are converging to some underlying true rating, or whether we need more samples or replicas.

Figure 10 and 11 show that the average scores have roughly converged around 150 samples per configuration, or around 15 replicas per sample. The two figures demonstrate this for nucleus sampling, and this holds true for human evaluations of all sampling algorithms.

E Additional Model-Generated Samples

Table 2 shows some additional samples from each of the sampling algorithms described in the paper. Similarly, we have chosen hyperparameters for each sampling method that yields a similar diversity (measured by self-BLEU) to the top- k configuration where $K = 15$. We observe that all sampling

Sampling	Conditional Samples
Existing Sampling Algorithms	
<i>Top-K</i> ($K = 15$)	<i>as the rest of his denver broncos teammates</i> prepared for the game against denver, jay kasey could not help but think of his teammates and friends who worked hard in preparation for that night’s game.
<i>Nucleus</i> ($P = 0.65$)	<i>as the rest of his denver broncos teammates</i> slumped and buried themselves in their work, broncos quarterback leon johnson moved to the locker room monday and called his parents.
<i>Temperature</i> ($T = 0.7$)	<i>as the rest of his denver broncos teammates</i> gathered in an auditorium to watch more stretching drills, ben holtz gave an emotional speech : we’re running out of time to win a championship ring.
Property-satisfying Sampling Algorithms	
<i>Random Top-K</i> ($R = 30$)	<i>as the rest of his denver broncos teammates</i> battled through their own stretch of the nfl playoffs, the quarterback began throwing the ball in the fourth quarter.
<i>Max Entropy</i> ($E = 2.75$)	<i>steven spielberg’s dreamworks movie studio</i> has agreed to pay \$ ## million to director john nichols (£ ## million, ###, a record in the studio circulation), the studio announced sunday..
Property-violating Sampling Algorithms	
<i>Random Mask</i> ($R = 0.75$)	<i>as the rest of his denver broncos teammates</i> connect with a player that the team didn’t expect to become a starter, quarterback james crosby speaks out about colin peterson’s passion for the game.
<i>Noised Top-K</i> ($K=20, W=5e-3$)	<i>as the rest of his denver broncos teammates</i> start making room for nerdy bundles or twigg pitchers, coach william perez might have to cut a big, bold note cut ready to console wife join them in iraq.
<i>Target Entropy</i> ($E = 2.5$)	<i>as the rest of his denver broncos teammates</i> scratched out their locker rooms, cleanDeath Yo Communities wander edge extingustretched cords429 Mohnegie wildfires.

Table 2: The samples conditioned on *as the rest of his denver broncos teammates*, and the hyperparameters for a given sampling algorithm. The poor quality spans are highlighted in red.

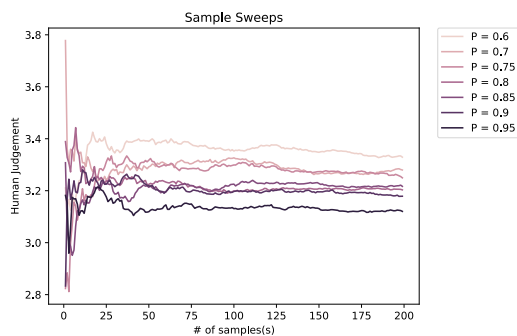


Figure 10: We see that we obtain a reasonable estimate of sample quality around 150 samples per configuration.

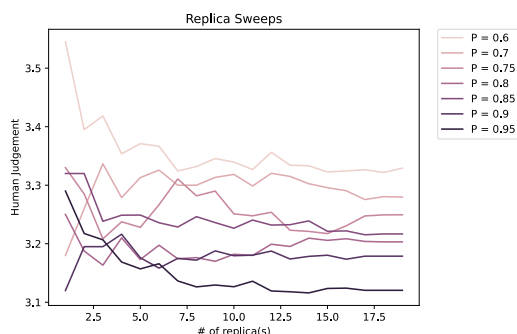


Figure 11: We see that we obtain a reasonable estimate of sample quality with around 15 ratings per sample.

algorithms except for noised top- k and target entropy, yield similar quality samples. For noised top- k and target entropy, we see that these samples tend to degenerate towards the end of the sentence,

indicating violation of the identified properties may possibly lead towards degraded performance.

F Human Evaluation with Self-BLEU as Diversity Metric

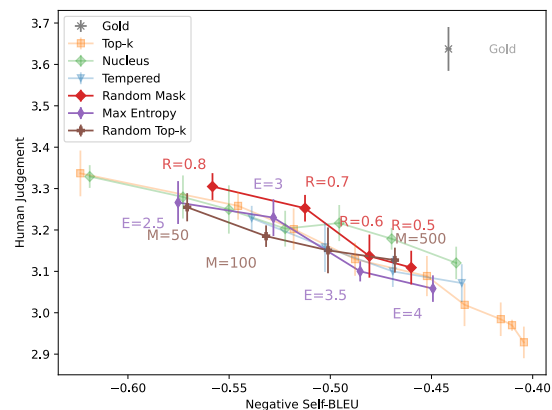


Figure 12: Using self-BLEU as a diversity metric provides similar conclusions as to using n-gram entropy.

Figures 1 and 5 measures diversity in terms of 3-gram entropy, while the rest of our work measures diversity in terms of self-BLEU. For completeness, we provide Figure 12 where self-BLEU is used for diversity metric. This figure demonstrates that similar trends can be observed using either 3-gram entropy or self-BLEU.