

# On Decoding Strategies for Neural Text Generators


Gian Wiher Clara Meister Ryan Cotterell

ETH Zürich, Switzerland

{gian.wiher, clara.meister, ryan.cotterell}@inf.ethz.ch

## Abstract

When generating text from probabilistic models, the chosen decoding strategy has a profound effect on the resulting text. Yet the properties elicited by various decoding strategies do not always transfer across natural language generation tasks. For example, while mode-seeking methods like beam search perform remarkably well for machine translation, they have been observed to lead to incoherent and repetitive text in story generation. Despite such observations, the effectiveness of decoding strategies is often assessed on only a single task. This work—in contrast—provides a comprehensive analysis of the interaction between language generation tasks and decoding strategies. Specifically, we measure changes in attributes of generated text as a function of both decoding strategy and task using human and automatic evaluation. Our results reveal both previously observed and novel findings. For example, the nature of the diversity–quality trade-off in language generation is very task-specific; the length bias often attributed to beam search is not constant across tasks.

 <https://github.com/gianwiher/decoding-NLG>

## 1 Introduction

Modern neural networks constitute an exciting new approach for the generation of natural language text. Much of the initial research into neural text generators went into designing different architectures (Sutskever et al., 2014; Rush et al., 2015; Serban et al., 2017). However, recent work has hinted that which **decoding strategy** (i.e., the method used to generate strings from the model) may be more important than the model architecture itself. For instance, a well replicated recent result is that, under a probabilistic neural text generator trained with the maximum-likelihood objective, the most probable string is often *not* human-like or high quality (Stahlberg and Byrne, 2019; Eikema and Aziz, 2020). In light of this finding, a plethora of decoding strategies have

been introduced in the literature, each claiming to generate more desirable text than competing approaches.

Lamentably, empirical studies of decoding strategies are typically evaluated on a *single* natural language generation task—without investigation into how performance may change *across* tasks—despite the fact that these tasks differ qualitatively across a large number of axes. These qualitative differences manifest quantitatively as well: For example, we can see in Figure 1 that high probability strings are favorable in some tasks, like machine translation (MT), while heavily disfavored in others, like story generation (SG). Consequently, we should not a priori expect a strategy that works well for one task to demonstrate the same performance in another. Indeed, several cases already show evidence of this: Beam search works remarkably well for machine translation but, outside of this context, has been observed to return dull text or degenerate text (Holtzman et al., 2020; DeLucia et al., 2021). This raises a natural fear that decoding strategies have been optimized for performance on a specific task, and the task-agnostic claims about the effectiveness of one decoding strategy over another are potentially ill-founded. A broader analysis of decoding strategies—both within and across tasks—is needed in order to fully understand the extent of such a problem.

Our work fills this gap, providing the first comprehensive comparison of decoding strategies across natural language generation tasks. Empirically, we compare strategy performance on several axes, taxonomizing methods into groups such as deterministic and stochastic, to understand the importance of various strategy attributes for quantifiable properties of text. In summary, our main findings include the following:

- Many previous empirical observations, among them the quality–diversity and quality–probability trade-offs (Ippolito et al., 2019;

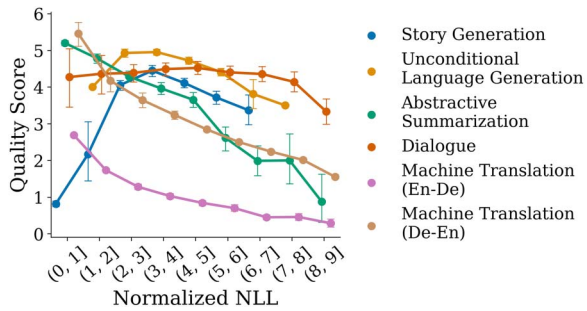


Figure 1: Quality-probability trade-off for different language generation tasks: story generation, unconditional language generation, abstractive summarization, dialogue, and machine translation. Notably, general trends in each curve differ drastically across tasks, despite training models with the same objective. See §4.3 for details on how quality scores are computed.

Zhang et al., 2021; Nadeem et al., 2020), manifest themselves in very task-specific ways. For example, our experiments reveal a distinct quality-diversity trade-off albeit only in a certain subset of tasks. This brings into question whether there is a single phenomenon under consideration or many distinct, but related, phenomena.

- A group-level analysis shows the first empirical evidence of a distinct divide in preference for stochastic versus deterministic strategies across tasks: All directed generation tasks appear to favor the latter, yet there is a notable trend in the strength of this preference—even the inverse is true for story generation.

We see these results as both a reference point for language generation practitioners, so that they can more confidently choose a decoding strategy that fits their needs, and as an indicator of potential strengths and weaknesses of today’s neural probabilistic language generators. We have reason to believe that there is a task-specific optimization happening in the literature whereby many of the proposed and (even celebrated) decoding strategies only outperform their competitors on specific tasks. Thus, our paper serves as a cautionary note about proper comparisons.

## 2 Probabilistic Language Generators

In this work, we consider models for language generation tasks that define a probability distribution over strings. More formally, these models are

probability distributions  $q$  over an output space  $\mathcal{Y}$ —(perhaps) conditioned on an input  $\mathbf{x}$ —where  $\mathcal{Y}$  is the set consisting of all possible strings that can be constructed from the vocabulary  $\mathcal{V}$ :

$$\mathcal{Y} \stackrel{\text{def}}{=} \{\text{BOS} \circ \mathbf{v} \circ \text{EOS} \mid \mathbf{v} \in \mathcal{V}^*\} \quad (1)$$

where BOS and EOS stand for special reserved beginning-of-sentence and end-of-sentence tokens, respectively;  $\mathcal{V}^*$  is the Kleene closure of  $\mathcal{V}$ .

Today’s language generators are typically parameterized by encoder–decoder architectures with attention mechanisms (Sutskever et al., 2014), notably the transformer (Vaswani et al., 2017), with trainable weights  $\theta$ . These models follow a local-normalization scheme, meaning that for all  $t > 0$ ,  $q(\cdot \mid \mathbf{y}_{<t})$  defines a probability distribution over  $\bar{\mathcal{V}} \stackrel{\text{def}}{=} \mathcal{V} \cup \{\text{EOS}\}$ . The probability of a sequence  $\mathbf{y} = \langle y_0, y_1, \dots \rangle$  can thus be factorized as follows:

$$q(\mathbf{y}) = \prod_{t=1}^{|\mathbf{y}|} q(y_t \mid \mathbf{y}_{<t}) \quad (2)$$

where  $\mathbf{y}_{<t} \stackrel{\text{def}}{=} \langle y_0, \dots, y_{t-1} \rangle$  and  $\mathbf{y}_{<1} = y_0 \stackrel{\text{def}}{=} \text{BOS}$ .

In order to learn the weights  $\theta$ , we minimize some loss function  $L(\theta; \mathcal{C})$ , defined in terms of a corpus  $\mathcal{C}$ . In theory, we want examples in  $\mathcal{C}$  to be assigned high probability. Accordingly, our loss is typically their negative log-likelihood under  $q$ .<sup>1</sup>

## 3 The Decoding Problem

We define the decoding problem as the search for a string  $\mathbf{y}^*$  according to a given model  $q$  and a set of decision rules. Given the probabilistic nature of most language generators, the natural choice for such a string would be the most probable sequence under the model

$$\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} \log q(\mathbf{y} \mid \mathbf{x}) \quad (3)$$

Solving the optimization problem in Eq. (3) is commonly referred to as maximum a posteriori (MAP) decoding. There are two main reasons

<sup>1</sup>For certain tasks, this loss is typically augmented with **label smoothing** (Szegedy et al., 2016) to combat overfitting. In short, a certain probability mass is discounted from the ground-truth token and redistributed uniformly across all the other tokens.

why in practice this direct optimization is not used when decoding: First, because of the exponentially large space  $\mathcal{Y}$  and the non-Markovian structure of commonly used neural generators, direct optimization is often computationally infeasible. Second, recent research has shown that the mode (i.e., the MAP solution  $\mathbf{y}^*$ ) is often not human-like or high quality text (Eikema and Aziz, 2020). For example, in the domain of MT, the most likely string under the model is often the empty string (Stahlberg and Byrne, 2019). For open-ended generation,<sup>2</sup> it has been observed that there is a positive correlation between likelihood and quality up to only a certain inflection point, after which the correlation becomes negative (Zhang et al., 2021). Thus in practice,  $\mathbf{y}^*$  is almost exclusively approximated using heuristic methods. An overview of such (commonly used) methods is presented below.

### 3.1 Deterministic Algorithms

**Greedy Search.** One approximation of  $\mathbf{y}^*$  is obtained by greedily choosing the most probable token at each decoding step  $t$ , that is, the following recursion is performed until the EOS symbol is chosen or some maximum time step  $T$  is reached:

$$\begin{aligned} y_0 &= \text{BOS} \\ y_t &= \operatorname{argmax}_{y \in \bar{\mathcal{V}}} \log q(y \mid \mathbf{x}, \mathbf{y}_{<t}) \quad (\text{for } t > 0) \end{aligned} \quad (4)$$

Note that there is no formal guarantee that greedy decoding will return the global optimum of the decoding objective since decisions are only locally optimal.

**Beam Search.** Beam search is a simple extension of greedy search. Rather than considering only the highest probability continuation of our string at each step, we keep the  $K \in \mathbb{Z}_+$  highest probability paths, where the hyperparameter  $K$  is referred to as the beam:

$$\begin{aligned} Y_0 &= \{\text{BOS}\} \\ Y_t &= \operatorname{argmax}_{\substack{Y'_t \subseteq \mathcal{B}_t, \\ |Y'_t|=K}} \sum_{\mathbf{y} \in Y'_t} \log q(\mathbf{y} \mid \mathbf{x}) \quad (\text{for } t > 0) \end{aligned} \quad (5)$$

<sup>2</sup>We define *directed* generation tasks as involving a strong relationship between input and output (e.g., as in MT); for *open-ended* tasks, input contexts only pose a soft constraint on the output space, i.e., there is a considerable degree of freedom in what is a plausible output (e.g., in dialogue or story generation).

where for  $t > 0$

$$\mathcal{B}_t = \left\{ \mathbf{y}_{t-1} \circ y \mid y \in \bar{\mathcal{V}} \text{ and } \mathbf{y}_{t-1} \in Y_{t-1} \right\} \quad (6)$$

is our beam, consisting of all possible extensions of  $\mathbf{y} \in Y_{t-1}$ . As with greedy decoding, the recursion is performed until all strings end in the EOS symbol or some maximum time step  $T$  is reached. The highest scoring string  $\mathbf{y}^*$  is then chosen from this final set  $Y_T$ .

Beyond the log-probability, other scoring functions have been proposed as modifications to the vanilla beam search algorithm. For example, Vijayakumar et al. (2018) propose **diverse beam search** (DBS) to address the issue of the lack of diversity within the set of returned strings. The algorithm further splits the beam into several subgroups and adds an inner iteration at each time step to maximize inter-group diversity. We refer the reader to the original work for the full algorithm.

### 3.2 Stochastic Algorithms

**Ancestral Sampling.** Instead of approximating  $\mathbf{y}^*$ , one can obtain generations by sampling  $\mathbf{y} \sim q(\cdot \mid \mathbf{x})$ . Due to the local normalization scheme of the models that we consider, this can be achieved simply by setting  $y_0 = \text{BOS}$  and then drawing each  $y_t \sim q(\cdot \mid \mathbf{x}, \mathbf{y}_{<t})$  until EOS is sampled.

**Top- $k$  Sampling.** Perhaps due to the “unreliable tail” of the distribution (Holtzman et al., 2020)—that is, the subset of  $\bar{\mathcal{V}}$  that are unrealistic extensions of a string but are necessarily assigned probability mass due to the non-sparse nature of the softmax transformation—sampling directly from  $q(\cdot \mid \mathbf{x})$  can lead to text that is incoherent and sometimes even unrelated to the subject (Fan et al., 2018). One way to overcome this issue is to limit the sampling space to the top- $k$  most likely tokens in each decoding step. Prior to sampling, the distribution over  $\bar{\mathcal{V}}$  is recomputed: Let  $Z_k(\mathbf{x}, \mathbf{y}_{<t}) \stackrel{\text{def}}{=} \sum_{y \in \bar{\mathcal{V}}^{(k)}} q(y \mid \mathbf{x}, \mathbf{y}_{<t})$  where  $\bar{\mathcal{V}}^{(k)} \subseteq \bar{\mathcal{V}}$  is defined to be the set of the  $k$  most likely tokens. The truncated distribution is given by:

$$\pi(y \mid \mathbf{x}, \mathbf{y}_{<t}) = \begin{cases} \frac{q(y \mid \mathbf{x}, \mathbf{y}_{<t})}{Z_k(\mathbf{x}, \mathbf{y}_{<t})} & \text{if } y \in \bar{\mathcal{V}}^{(k)} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Task	$\mathcal{X}$	$\mathcal{Y}_{\text{out}}$	$q$	Dataset
Machine Translation (MT)	sequence in source language	sequence in target language	FAIR’s WMT19 submission	NEWSTEST2019
Abstractive Summarization (AS)	news article	summary	BART	CNN/DAILYMAIL
Dialogue (Diag)	conversation history	response	DIALOGPT	DIALOGPT
Story Generation (SG)	short prompt	related story	GPT-2 (small and medium)	WRITINGPROMPTS
Unconditional Generation (ULG)	empty sequence ( $\langle\langle\text{bos}\rangle\rangle$ )	plausible natural language strings	GPT-2 (small and medium)	WIKITEXT-103

Table 1: Overview of the tasks considered in this work. Examples given for  $\mathcal{X}$  and  $\mathcal{Y}_{\text{out}}$  are the intended input and output, respectively. Models  $q$  are evaluated on the test set of the specified dataset. Note that we fine-tune the GPT-2 models on the specified dataset, while other models are loaded from checkpoints provided by the Hugging Face framework (Wolf et al., 2020).

**Nucleus (Top- $p$ ) Sampling.** Rather than always considering a fixed size set, nucleus sampling dynamically adjusts the number of tokens under consideration based on the spread of the probability distribution at each generation step. Formally, nucleus sampling (Holtzman et al., 2020) considers the smallest subset of tokens whose cumulative probability mass exceeds a chosen threshold  $p$ . For generation step  $t$  and  $p \in (0, 1]$ , the top- $p$  vocabulary  $\bar{\mathcal{V}}^{(p)} \subseteq \bar{\mathcal{V}}$  is defined as the smallest set such that

$$\sum_{y \in \bar{\mathcal{V}}^{(p)}} q(y \mid \mathbf{x}, \mathbf{y}_{<t}) \geq p \quad (8)$$

The truncated distribution is computed similar to Eq. (7) with  $Z_p(\mathbf{x}, \mathbf{y}_{<t}) \stackrel{\text{def}}{=} \sum_{y \in \bar{\mathcal{V}}^{(p)}} q(y \mid \mathbf{x}, \mathbf{y}_{<t})$ .

**Bayes Minimum Risk (MBR).** Under probabilistic language generators, probability mass is often spread over a large set of likely candidates without clear preference (Ott et al., 2018). However, this set of likely strings should not be arbitrary when  $q$  is good. Rather, these strings should capture the statistics of training data well, containing a number of potentially good solutions (Eikema and Aziz, 2020). This motivates a decision rule that exploits all available information in this set. Let  $u: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  be a utility function that evaluates a string  $\mathbf{y}$  against reference  $\hat{\mathbf{y}}$ . According to statistical decision theory (Bickel and Doksum, 1977), the optimal decision  $\mathbf{y}^*$  is the one that minimizes expected risk

$$\mathbf{y}^* = \operatorname{argmin}_{\mathbf{y} \in \mathcal{Y}} \mathbb{E}_{q(\hat{\mathbf{y}}|\mathbf{x})}[-u(\mathbf{y}, \hat{\mathbf{y}})] \quad (9)$$

Like MAP, it is generally computationally infeasible to solve the MBR objective exactly given the size of  $\mathcal{Y}$ . In practice, one can obtain an unbiased estimate of the expected risk via Monte Carlo

(MC) methods, limiting the search space for the maximization problem to the set sampled during this estimation procedure.

## 4 Experimental Setup

The strategies presented in Section 3 are compared across a variety of NLG tasks covering open-ended as well as directed generations tasks. We define a task more formally as a triple  $(\mathcal{X}, \mathcal{Y}_{\text{out}}, q)$  where  $\mathcal{X}$  denotes the input space,  $\mathcal{Y}_{\text{out}} \subseteq \mathcal{Y}$  the output space<sup>3</sup> and  $q$  a model that defines a probability distribution over  $\mathcal{Y}_{\text{out}}$  for every input  $\mathbf{x} \in \mathcal{X}$ . A high-level overview of these tasks (and the respective datasets used) can be found in Table 1. We use solely transformer-based models, all state-of-the-art for their respective tasks (Ng et al., 2019; Lewis et al., 2020; Zhang et al., 2020; Radford et al., 2019). We use open-sourced versions of models for reproducibility.

**Decoding Strategy Settings.** Most of the decoding algorithms specified in the previous section depend on certain parameters. For all our experiments we use the following settings:

- We consider **beam search** with beam sizes  $K = 5$  and  $K = 10$ , and **DBS** with Hamming distance as a dissimilarity function,  $\lambda = 0.7$  and  $G = K = 5$ . The choice of dissimilarity function and hyperparameters is based on the recommendations from the original work. When we only want to return one string, we select the hypothesis with the highest score according to  $\log q$ .
- For **top- $k$**  sampling, we set  $k = 30$  and for **top- $p$**  sampling, we set  $p = 0.85$  based on

<sup>3</sup>Note that formally the input and output spaces only differ by the model-specific vocabularies and maximum generation length  $l \in \mathbb{Z}_+$ .

Quality Metrics	
Automatic	
BLEU	Corpus-level metric originally developed to assess translation quality of MT systems (Papineni et al., 2002). Produces a score between 0 and 1 based on modified $n$ -gram precision. We use the SACREBLEU (Post, 2018) framework.
METEOR	Metric based on the harmonic mean of unigram precision and recall. Originally developed to evaluate MT. We use version 1.5 of the implementation from Denkowski and Lavie (2014).
COMET	Neural framework to train multilingual MT evaluation systems proposed by Rei et al. (2020). The nature of these metrics makes it only compatible with the MT task. We use a pretrained model checkpoint provided by the original work.
ROUGE	Recall-oriented set of metrics originally developed to assess the quality of automatically generated summaries (Lin, 2004). We report the ROUGE-L measure, which is based on longest common subsequences between candidate and reference.
BLEURT	Trained evaluation metric based on BERT (Devlin et al., 2019). Returns a score that indicates to what extent the candidate is grammatical and conveys the meaning of the reference (Sellam et al., 2020). We use a pretrained model checkpoint provided by the original work.
Human	
ADEQUACY	How well does the response/continuation fit in a given conversation history?
NATURALNESS	To what degree does the text seem to be a natural English text?
QUALITY	How high is the overall quality of the text?
ACCURACY	Given the context, is the text accurate?
FLUENCY	How fluent is the given text?
Diversity Metrics	
DIST- $n$	Number of distinct $n$ -grams divided by the total number of $n$ -grams (Li et al., 2016).
ENT- $n$	The fact that infrequent $n$ -gram contribute more to diversity than frequent ones is not taken into account by dist- $n$ . This limitation is addressed by the ent- $n$ metric first proposed by Zhang et al. (2018) which reflects how uniform the empirical $n$ -gram distribution is for a given sentence.
$n$ -GRAM DIV.	Average over dist- $n$ measures for different values of $n$ . We calculate the average over $n \in \{1, \dots, 5\}$ .
SELF-BLEU	Average BLEU score across strings when using all other strings in set as references (Zhu et al., 2018).
REPETITION	If a phrase (minimum length of 2) is repeated at least three times until the end of the generation, it is labeled as a repetition. This definition of a repetition is taken from Holtzman et al. (2020)

Table 2: List of metrics considered in this work. For human evaluation metrics, prompt shown is provided to raters.

experiments in DeLucia et al. (2021) that suggest a parameter range  $p \in [0.7, 0.9]$ .

- For **MBR**,<sup>4</sup> we obtain 30 to 32 ancestral samples<sup>5</sup> to approximate the expected risk in Eq. (9) using MC. The candidate sequences,

for which we all calculate the expected risk, consists of the ancestral samples used for the MC approximations together with sequences obtained from the other decoders. The metric BEER (Stanojević and Sima'an, 2014) is used as the utility function  $u$ .

<sup>4</sup>We use code provided at [github.com/RoxoT/mbr-nmt](https://github.com/RoxoT/mbr-nmt).

<sup>5</sup>To speed up the generation process, samples are generated in batches. Depending on the memory requirements of

the different models, the batch size differs across tasks, thus creating small differences in the number of samples acquired.

Task	ADEQUACY	QUALITY	FLUENCY	NATURALNESS	ACCURACY
Diag	✓			✓	
AS		✓			✓
SG			✓	✓	
ULG			✓	✓	

Table 3: Criteria used for each task in human evals.

#### 4.1 Metrics

We use a number of different metrics to compare text across decoding strategies. An overview of all metrics can be found in Table 2. Note that we roughly divide the set of metrics into two categories: *diversity* metrics and *quality* metrics. Intuitively, we may expect that the two criteria are not always of equal importance. For example, in MT an accurate, high quality translation of the input is often more highly valued than generating engaging or stylized language or a wider range of diverse outputs. On the other hand, a conversational agent that is able to talk about a diverse range of topics is likely highly preferred to one that repeats the safest phrases over and over (Li and Jurafsky, 2016). In our subsequent experiments, we provide a quantitative analysis of this trade-off.

#### 4.2 Evaluation of Quality

For tasks where one has access to a ground truth reference, for example, MT, AS, and to some extent Diag, there are a variety of automatic metrics to evaluate quality. Most of these metrics are based on statistics of  $n$ -gram overlap between output and reference. This class of metrics has its limitations; consequently we also consider human judgments of text quality using criteria in Table 2. We use the *prolific* framework to obtain ratings from 5 different annotators on 200 examples per decoding strategy; criteria used for each of the tasks is given in Table 3. For each of the criteria, an 8-point Likert scale is used. We select the criteria based on which have been most commonly used to assess performance of text generators on a given task, as outlined by van der Lee et al. (2021), and describe them to the annotators as in Table 2. If a rater assigns high scores to multiple examples that do not fulfill the specified criteria at all, the rating is rejected and we obtain a fresh set of scores from a new rater. For SG, AS, and Diag the raters are first presented with a prompt/news article/dialogue history followed by the out-

puts of the different decoders and the reference in random order. For unconditional language generation we present the raters with generations and references in random order.<sup>6</sup>

#### 4.3 Evaluation of Diversity

Automatic metrics to measure lexical diversity of generated text are mostly based on statistics of  $n$ -gram counts; while lexical diversity is a narrow definition of diversity, it is the commonly employed one in language generation as diverse word choice is arguably a large factor for this characteristic. Note that lexical diversity can be measured at the string level, that is, within a given string  $\mathbf{y}$ , or across a set of strings  $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots\}$ . While we provide some results for the former set of metrics, we focus largely on the latter set, as often practitioners are more concerned with having a diverse set of generations *per* input. Specifically, we take measurements with respect to sets decoded by each strategy, namely, the size  $K$  set decoded by beam search or  $K$  items generated according to a specific stochastic scheme.<sup>7</sup> For the stochastic decoders, we set  $K = 10$ . For each input  $\mathbf{x} \in \mathcal{X}$  we thus obtain a set of outputs, denoted by  $\mathcal{S}_{\mathbf{x}}$ , per decoder over which we calculate various metrics, such as self-BLEU or  $n$ -gram diversity. Self-BLEU is calculated on a per-string basis as the average of BLEU scores when setting one of the generations  $\mathbf{y}^{(i)} \in \mathcal{S}_{\mathbf{x}}$  as the hypothesis and all other strings in  $\mathcal{S}_{\mathbf{x}} \setminus \{\mathbf{y}^{(i)}\}$  as references. To calculate dist- $n$ , ent- $n$ , and  $n$ -gram diversity metrics for a set of generations, we concatenate all outputs and perform calculations as described in Table 2. For ULG, where we only have one input  $\mathbf{x}$ , we instead calculate scores over random (disjoint) subsets of size  $K = 10$ .

### 5 Results

#### 5.1 Quality

Human evaluations are aggregated across raters, using the median value for each string. Results are displayed in Figure 2. According to human raters, sampling directly from the model yields

<sup>6</sup>We omit human annotations for MT because it has been observed that there is no significant gain over the automatic metrics when using crowd workers due to large variations in evaluation (Freitag et al., 2021).

<sup>7</sup>Because greedy and MBR decoding are methods that only return a single string, they are not considered in the latter set of metrics.

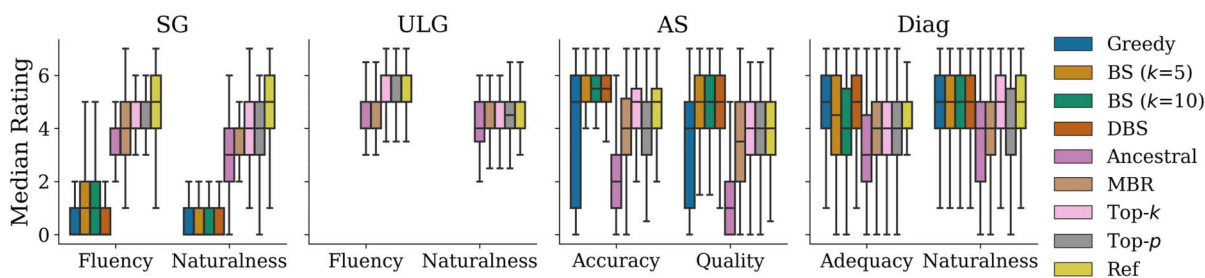


Figure 2: Median human evaluation ratings.

	Dialogue			Summarization			MT (De-En)			MT (En-De)		
	HUMAN	BLEU	ROUGE-L	HUMAN	BLEU	ROUGE-L	BLEU	METEOR	COMET	BLEU	METEOR	COMET
Greedy	4.660	0.661	<b>9.072</b> <sub>(4)</sub>	3.671	16.560	28.027	40.083	42.444	0.548	42.072	59.174	0.613
BS ( $k=5$ )	4.495	<b>0.758</b> <sub>(3)</sub>	8.796	<b>5.235</b> <sub>(5)</sub>	17.197	31.138	41.049	43.005	<b>0.561</b> <sub>(3)</sub>	<b>42.746</b> <sub>(5)</sub>	<b>59.602</b> <sub>(3)</sub>	0.622
BS ( $k=10$ )	4.456	0.746	8.331	5.180	16.726	30.650	<b>41.211</b> <sub>(5)</sub>	<b>43.101</b> <sub>(3)</sub>	0.560	42.680	59.583	<b>0.622</b> <sub>(3)</sub>
DBS	<b>4.689</b> <sub>(3)</sub>	0.436	8.708	5.122	<b>18.141</b> <sub>(5)</sub>	<b>31.487</b> <sub>(6)</sub>	39.770	42.254	0.538	41.702	58.793	0.611
MBR	3.815	0.510	8.469	3.709	10.771	25.120	40.811	42.952	0.547	42.370	59.241	0.605
Ancestral	3.329	0.196	5.408	1.825	5.390	17.985	17.402	27.425	-0.520	15.595	35.722	-0.832
Top- $k$	4.234	0.308	6.961	4.276	11.644	25.961	27.574	35.651	0.376	27.091	47.839	0.458
top- $p$	3.914	0.308	6.331	3.976	11.785	25.505	29.397	36.704	0.382	29.998	49.778	0.481

Table 4: Corpus-level quality metrics for Diag, AS, and MT. For Diag and AS the human score is calculated by taking the mean over the two criteria upon which the text is rated.

text with the lowest quality metrics across all tasks: The clear exception is for SG, where we observe that mode-seeking strategies lead to degenerate text (further discussion in §5.4). In general, for the directed generation tasks (AS and Diag), beam search variants perform the best, even outperforming human generated references. Interestingly, despite its limited exploration of the search space, greedy decoding generates texts on par with beam search methods for Diag.

On the other hand, the results of stochastic methods are more nuanced: Although top- $p$  and top- $k$  decoding generate more highly rated texts than ancestral sampling, they often fail to reach quality levels of the beam search based methods. MBR decoding, which as a decoding strategy perhaps falls somewhere between the classes of deterministic and stochastic, likewise performs somewhere between these classes in terms of quality metrics. Overwhelmingly, trends in performance are much more distinct when analyzing strategies as stochastic vs. deterministic, rather than individually, suggesting that small algorithmic differences in decoding strategies may not be as critical as prior work has made seem.

We present automatic quality evaluation metrics for directed generation tasks in Table 4—the number in brackets in Table 4 shows how many of the decoders performed *significantly* worse than the best one in terms of the respective met-

ric, as determined by an example-level permutation test. We use a significance level of 0.01; the resulting  $p$ -values were corrected for multiple testing using a Bonferroni correction. We observe similar trends as with our human evaluations: Beam search methods perform best, followed by top- $p$  and top- $k$  sampling, with ancestral sampling performing worst. Despite mixed results in Figure 2, MBR decoding yields competitive results in terms of automatic evaluation metrics, even matching the performance of beam search; this is perhaps not surprising given the poor correlation between human and automatic evaluation that is frequently observed in language generation. On Diag, we only observe a significant difference in performance between the best decoder and the worst 3, respectively, worst 4 decoders. Similarly, on MT, we observe that except for the BLEU metric, only a significant difference between the best and the worst 3 decoders is present. On the other hand, we have that for the AS the best performing decoder significantly outperforms any other decoder except the other beam search methods. This contrasts the observation for Diag and MT where the mode seeking decoders seem all to perform equal.

## 5.2 Diversity

We report diversity metrics for different strategies and tasks in Figure 3. Points are connected to



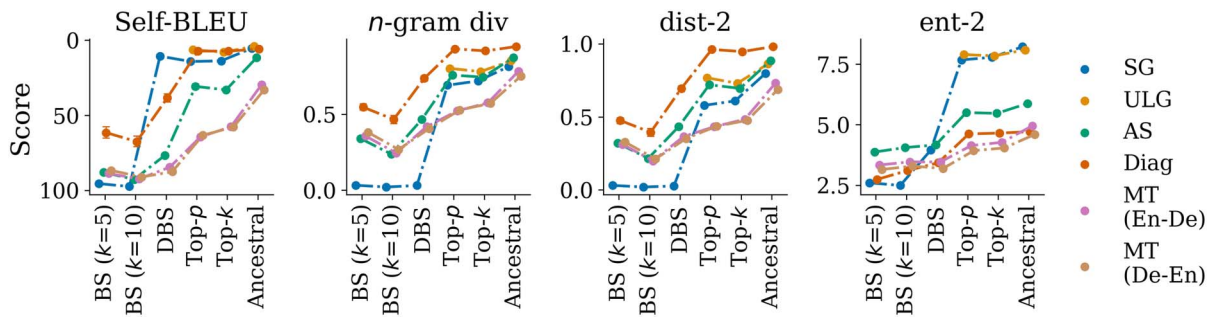


Figure 3: Diversity metrics calculated at the set level. For ULG, the metrics are calculated for randomly chosen (disjoint) subsets of all generations. Note that low self-BLEU indicates high diversity.

better illustrate general trends across diversity metrics, not due to a quantitative relationship between the metrics themselves. We see that in general, the trends for a given task are quite consistent across diversity metrics, that is, lines of the same color follow the same trend across facets. On the other hand, trends *across* tasks are not as similar. For example, the gap in diversity between deterministic and stochastic methods is much more exacerbated in SG than MT.

Across tasks, ancestral sampling consistently produces the most diverse outputs. Limiting the search space, as in top- $k$  and top- $p$  sampling, leads to a drop in diversity compared to pure sampling; notably, this drop appears to be much more significant for directed generations tasks. Interestingly, introducing a diversity promoting term, as in DBS, increases diversity with respect to beam-based decoding algorithms, but still leads to substantially less diverse strings than stochastic methods.

At the task-level, responses for Diag seem to be more inherently diverse than for other tasks. Even methods known for producing repetitive sets (e.g., beam search) generate a relatively diverse set of solutions. This suggests that even though diverse options are often desired in Diag, we may not need to explicitly optimize for them via the chosen decoding strategy. On the other hand, diversity in SG is quite sensitive to the chosen decoding strategy, displaying drastic differences.

### 5.3 Quantitative Trade-offs in NLG Tasks

We provide an analysis of the importance of different metrics for each of the language generation tasks, looking specifically at their relationships with perceived quality.

**The Probability–Quality Relationship.** Natural language generation is performed almost solely using probabilistic models. While ideally, we would like high quality text to be assigned high probability (and vice versa), we see that in practice this is not always the case (Cohen and Beck, 2019; Stahlberg and Byrne, 2019; Holtzman et al., 2020; Zhang et al., 2021; DeLucia et al., 2021). The trends observed in Figure 1 reveal that while high probability is often a determinant of quality in directed generation tasks, such as MT and AS,<sup>8</sup> there is a negative correlation between quality and probability in SG and ULG at least up until a certain inflection point. Such relationships have been a main motivation behind research into new decoding strategies (e.g., Li et al., 2016; Shao et al., 2017; Holtzman et al., 2020). The quality scores in Figure 1 are obtained by taking the mean over human ratings. For MT, sentence-BLEU is used.

This relationship also manifests in the divide in performance between deterministic strategies—all of which to some extent are mode-seeking—and stochastic strategies. Naturally, the deterministic decoding strategies we consider produce (on average) higher probability strings, as probability is part of their decoding objectives. Figure 6 shows that when compared to ancestral samples, most beam search generations are more strongly associated with higher (length) normalized log-likelihood than the output of the sampling based decoders. Thus, we might expect the results observed in Figure 1 to appear in a comparison of deterministic and stochastic strategies. We rank

<sup>8</sup>As computational constraints make it difficult (if not infeasible) to decode the highest probability string from neural models, we do not observe behavior at the extreme end of Figure 1, which other works have observed to produce poor quality text.



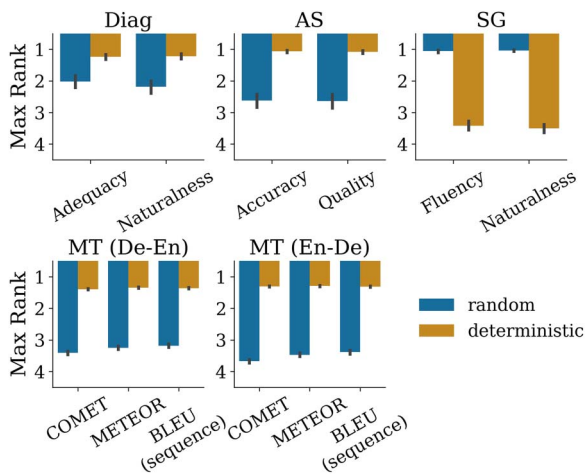


Figure 4: Highest ranks achieved by stochastic vs. deterministic strategies on each input; a rank of 1 means a generation from the respective group of decoding strategies was ranked 1st among all generations. We omit ULG since only stochastic strategies are considered for this task. Note that the lowest possible rank for a deterministic strategy is 4 and for a stochastic strategy is 5.

strategies within a task according to human ratings when available and calculate the highest rank obtained by each of the two groups. More specifically, for each input, we order generations according to their median human rating. Ranks are then assigned to each decoding strategy according to this ordering (lower is better). We then look at the highest rank achieved by the two subsets of decoding strategies. From Figure 4 we can see a distinct divide in preference for deterministic vs. stochastic strategies across tasks: All directed generation tasks appear to favor mode-seeking strategies. Yet there is a notable trend in the strength of this preference. As we might intuitively expect, we see an upward trend in the difference in rankings of mode-seeking vs. stochastic decoding methods as a task becomes more semantically constrained. At one end of the spectrum, in SG, we observe that in nearly all cases, the most highly ranked output from a deterministic strategy is still ranked below the worst of the stochastic strategies,<sup>9</sup> indicating the ill-suitedness of mode-seeking strategies for such tasks. The opposite is true of MT at the other end of the spectrum.

<sup>9</sup>This must be the case since the average maximum ranking for mode-seeking methods is almost 4.

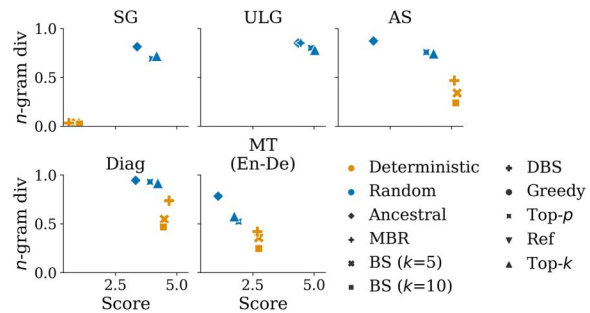


Figure 5: The relationship between diversity ( $n$ -gram div) and quality (median human rating) across language generation tasks. Results are qualitatively the same when using other diversity metrics, as we might expect given the results in Figure 3.

**The Diversity–Quality Relationship.** Here we investigate how diversity—as quantified by metrics in §4.1—relates to quality in a given task. Note that the probability–quality relationship has previously been attributed to a *trade-off* between diversity and quality (Zhang et al., 2021; Nadeem et al., 2020), albeit only in the investigation of a small subset of language generation tasks. However, we see in Figure 5 that the relationship between diversity and probability is not so easily defined: it changes quite drastically across tasks.

Specifically, Figure 5 shows there is indeed a trade-off for the two quantities in AS and MT, yet there appears to be an *interdependence* for open-ended generation tasks. Notably, Diag appears to fall outside of this paradigm, which perhaps challenges its definition as a directed generation task. In conjunction with other results (e.g., Figure 6), the trends shown in Figure 5 suggest that within directed tasks, Diag falls closer to open-ended generation tasks on the task spectrum. We further see that stochastic and deterministic methods are distinctly divided along the diversity–quality trend in each task; although this result is perhaps to some extent expected, the separating line is surprisingly sharp in all cases.

## 5.4 Eliciting Metrics

We now look at the ability of different decoding strategies to elicit the qualitative metrics described in §4.1, the quantitative properties studied in §5.3, as well as certain undesirable attributes of text. Through this analysis, we hope to ascertain how the effectiveness of different decoding strategies

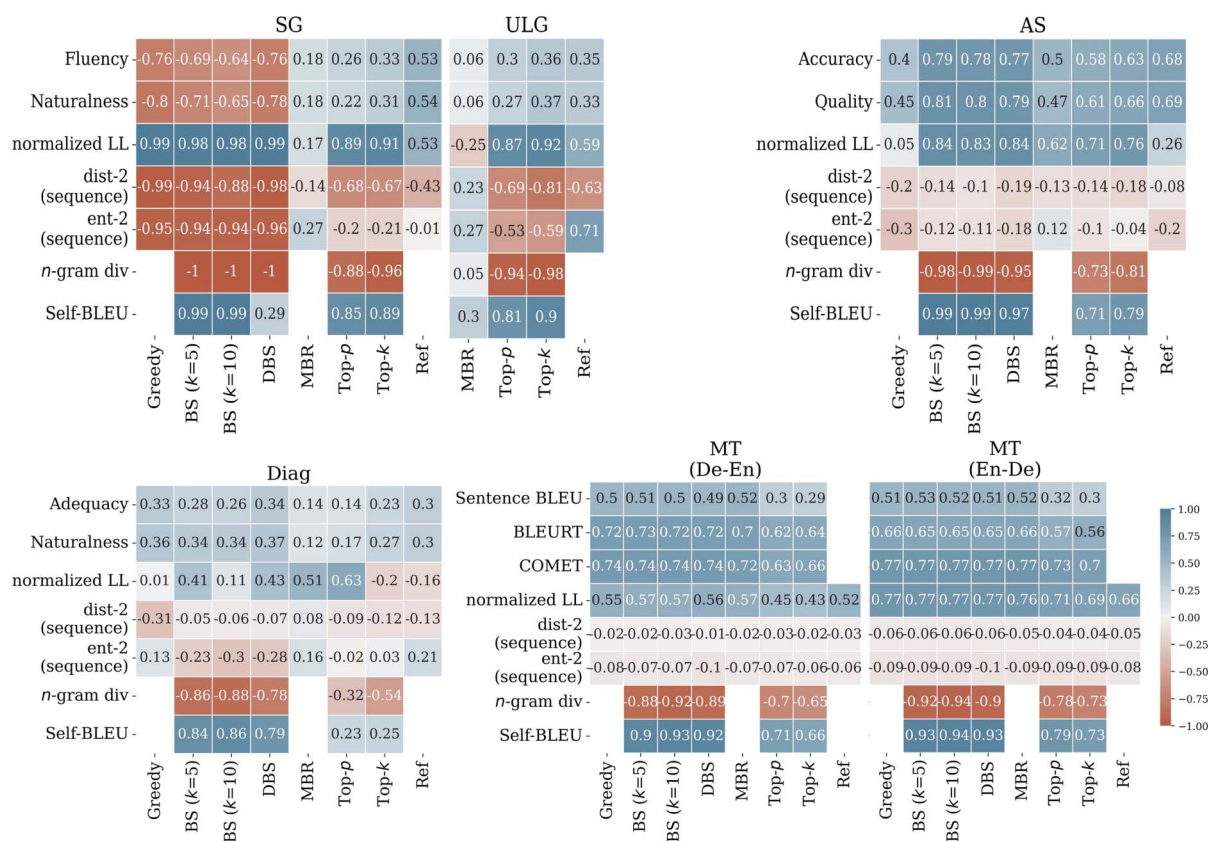


Figure 6: Correlations between quality metrics and other quantitative attributes of text with different decoding schemes separated by task.

generalizes across tasks, and which—if any—more general claims can be made about these strategies.

Figure 6 shows how different decoders correlate with various metrics, using ancestral samples as a baseline.<sup>10</sup> Our first take-away is that these correlation plots differ notably across tasks, which further demonstrates the sensitivity of the performance of decoding strategies to the task at hand. Among these differences though, we observe certain trends that provide insights into how decoders’ abilities to generate certain types of texts transfers across tasks. For example, the performance of decoders *within* the subsets of directed and open-ended generation tasks is reasonably consistent. We first discuss more specific trends with respect to *quality* metrics.

**Quality Metrics.** We first note that there is no single decoding method that consistently corre-

<sup>10</sup>Ancestral samples give us an unbiased sample of the type of text that is assigned probability mass by our model, thus making it a good baseline for observing the effects of decoding strategies.

lates most strongly with high-quality text, which heeds further warnings against more general claims made about decoder performance. Perhaps the most distinct result when looking at decoders’ correlations with quality metrics is the difference in correlations for mode-seeking methods between open-ended and directed generation tasks. Here we see that on the directed tasks, the use of mode-seeking methods appears to correlate highly with quality, with no substantial differences among this class of methods even when, for example, also optimizing for intra-set diversity (as in DBS). Interestingly, the strengths of the correlations shown by stochastic methods are much more consistent across all tasks than the mode-seeking methods. While in general, decoder performance with respect to quality metrics is relatively consistent for directed generation tasks, there are exceptions to this consistency: MBR correlates well with quality metrics for MT, but underperforms in comparison to other decoders for both AS and Diag. On AS, greedy search tends to lead to poorer quality text than top- $p$  and top- $k$  sampling where, for the other directed tasks,

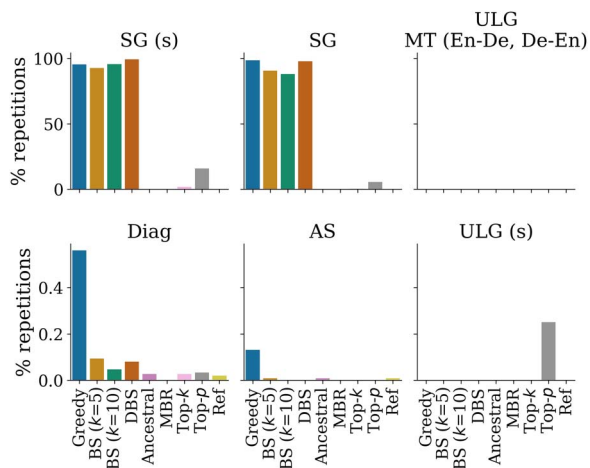


Figure 7: Fraction of generations that degenerate into repetition (see Table 2 for definition). Note the different scales for the different tasks.

all mode-seeking methods provide generations of higher quality.

**Diversity Metrics.** In comparison to quality, we observe that behavior of different decoders changes less with respect to diversity. For diversity metrics calculated over a set of generations, ancestral sampling consistently generates the most diverse text (as demonstrated by the negative  $n$ -gram diversity/positive self-BLEU correlations shown by all decoders). This is true even in comparison to DBS, which optimizes for intra-set diversity.<sup>11</sup> Mode-seeking decoding strategies consistently have a stronger negative correlation with set-level diversity metrics (e.g., self-BLEU) than their stochastic counterparts. This difference is more pronounced on certain tasks: For example, both Figure 6 and Figure 3 show a bigger jump in diversity scores between DBS and top- $p$  sampling on SG compared to MT or AS. Interestingly, there is little consistency across tasks in terms of sequence-level string diversity.

**Repetitions.** Probabilistic language generators are known to occasionally produce text with degenerate qualities (Dinan et al., 2020; Holtzman et al., 2020; Welleck et al., 2020b). One common form of degenerate behavior is repetitions, where generation falls into a loop of repeating the same phrase until the decoding algorithm terminates.

<sup>11</sup>Although in general, DBS seems to be relatively effective at optimizing for intra-set diversity in comparison to other decoders, even achieving low self-BLEU on SG despite also causing degeneration, as shown in Figure 7.

	Story Gen. (small)		Story Gen. (medium)	
	% repetition	ppl	% repetition	ppl
Greedy	95.67	1.07	98.47	1.09
BS ( $k = 5$ )	92.58	1.11	90.70	1.11
BS ( $k = 10$ )	95.67	1.11	88.01	1.11
DBS	99.25	1.05	97.75	1.05
MBR	0.20	27.19	0	28.46
Ancestral	0.23	30.43	0.13	32.98
Top- $k$	1.97	7.10	0.53	7.38
Top- $p$	15.87	5.52	5.65	6.33
Reference	0	23.83	0	19.28

Table 5: Perplexities and repetition count for different strategies on the SG task. Mode-seeking strategies are able to produce text with very low perplexity but these generations almost always degenerate into repetitions.

Here we analyze the fraction of times this behavior occurs for different strategies; results can be found in Figure 7. On the SG task, we observe a substantial amount of text degeneration for mode-seeking strategies; this holds true for both small (s) and medium variants of GPT-2. Across both open-ended tasks, the only stochastic decoding scheme that appears to elicit this degenerate behavior is top- $p$  sampling; Although only a small percentage of samples, it is responsible for all of the degenerate behavior observed for the ULG task. Notably, for all tasks besides SG, we see repetitive behavior in less than 1% of generations. The exact repetition counts together with the perplexity of the generated texts for SG are shown in Table 5.

**Length.** We further investigate how different decoding strategies affect the length of generated text. Length biases have frequently been observed in language generation tasks (Murray and Chiang, 2018; Welleck et al., 2020a), both for shorter and longer strings. In this experiment, we hope to observe how much the decoding scheme can be held responsible for these biases. We report results in Figure 8 and Figure 9. For MT, all strategies manage to generate strings of lengths similar to the reference with the exception of ancestral sampling, which produces slightly longer strings. Interestingly, there are no consistent trends for beam search variants across the other directed generation tasks; rather, trends seem to be inverted for Diag and AS.

We see large variation in the length of generated strings for the SG task, especially among

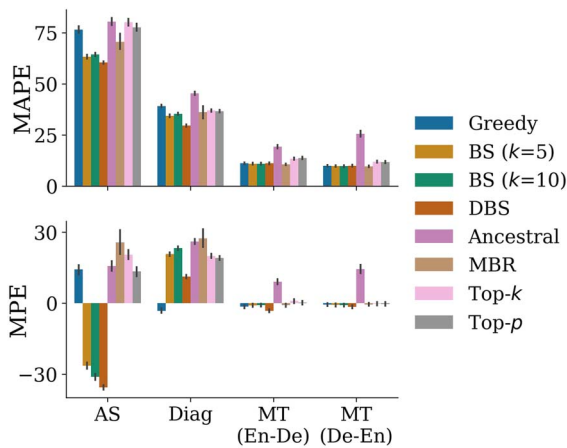


Figure 8: Differences between lengths of generated texts to reference strings. MAPE denotes the mean absolute percentage difference between reference lengths and the lengths of generated texts. MPE denotes the mean percentage error, where we do not take the absolute value of the difference in lengths, in order to get a sense of whether generated strings are (on average) longer or shorter than the reference.

mode-seeking strategies; for example, standard beam search produces rather short strings while DBS and greedy decoding produce inordinately long strings. For the unconditional language generation task, we observe no big differences in generated sequence length among stochastic methods. Collectively, these results tell us that the previously observed length biases are task–decoder specific, rather than purely decoder specific.

## 6 Discussion

When constructing a text generation pipeline, the choice of decoding strategy has a large effect on various aspects of the resulting text. Yet when making this choice for a specific language generation task, practitioners are currently limited to either basing their decision on non-comprehensive analyses, using expensive human annotations or even resorting to guesswork. There are potential pitfalls in these practices: As evidenced by various results in this work, certain properties of decoding schemes—especially quality—do not transfer across tasks. This work aims to provide guidance for practitioners in the choice of decoding strategies, revealing their strengths and weaknesses with respect to individual tasks while also giving insights into whether one can expect these properties to transfer to tasks outside of this study.

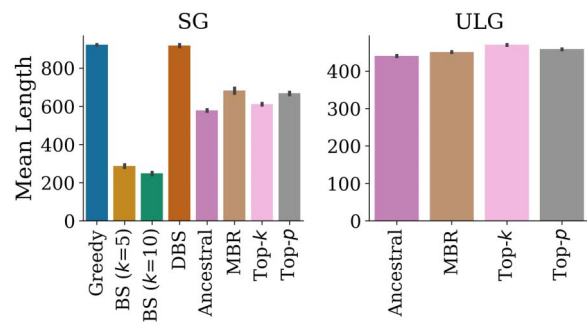


Figure 9: Mean lengths of generated text for open-ended tasks. Results are displayed from models based on the medium-sized version of GPT-2; we omit results for the small version, which were ostensibly the same.

While all of the takeaways from this work cannot be summarized in a few lines, we highlight some key observations below.

The relationships and trade-offs between certain properties of text changes notably from one task to another. For example, as depicted in Figure 1, high-probability strings are typically also of high quality for MT while there is an almost inverse relationship between these attributes for SG. As shown in Figure 5, a quality–diversity trade-off exists for directed generation tasks whereas for open-ended generation tasks, the relationship is almost a co-dependence. These task-specific characteristics must be taken into account when both choosing and developing decoding strategies.

While decoder performance generally does not transfer faithfully across tasks, we can still identify some rules from our experiments that practitioners can use. For one, we see that on directed generation tasks, mode-seeking methods all perform competitively in terms of quality. Further, for stochastic decoders, we observe that restricting the sample space—as done in top- $p$  and top- $k$  decoding—greatly increases quality compared to ancestral sampling, albeit sacrificing some diversity. The ability of a decoder to elicit diversity in text—at least at the set-level—is perhaps the most consistent decoder quality across tasks. There are many other use-case specific insights that can be drawn from the results shown by figures and statistics in this work, which we hope serve as further guidance for practitioners.

It is worth noting that the behavior of decoders depends on their respective hyperparameters, for example,  $k$  or  $p$  in top- $k$  and top- $p$  sampling. This work does not perform a thorough search over hyperparameters, instead utilizing those most



widely used in order to optimize for the usefulness of our results to practitioners, who are likely to use similar default settings. While based on the results of other works, these choices should provide representative variants of the text generated according to the respective decoding strategy, this is a limitation of our work worth taking into consideration.

## 7 Conclusion

This work provides an extensive analysis of the effects of different decoding strategies on generated text across various language generation tasks. We show how different attributes of model-generated text change depending not just on decoding strategy, but also on the task at hand, using both human and automatic evaluations. Our results both confirm several prior observations, for example, a trade-off between diversity and quality metrics for specific NLG tasks, while also revealing a number of previously unobserved trends in language generation, both with respect to decoding strategies and the tasks themselves. A main take-away of these results is that decoding strategies are perhaps optimized for specific language generation tasks and that practitioners should take great care in basing their choice of decoding strategy off of results reported for alternate tasks. We release the evaluation framework and generations in the hopes that this type of analysis will be extended, for example, by ablating components of model or training strategies, in order to isolate which artifacts can be attributed to the nature of a specific generation task vs. design choices. We ultimately see this line of research as important for helping practitioners more confidently choose a decoding strategy that fits their needs without the use of valuable resources, for the further development of decoding strategies and for better understanding the shortcomings of probabilistic language generators.

## Ethical Concerns

We do not foresee any ethical concerns.

## Acknowledgments

We would like to thank Bryan Eikema for his valuable insights and many fruitful discussions.

## References

- Peter J. Bickel and Kjell A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*, 2 edition, volume 1.
- Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of the International Conference on Machine Learning*, volume 97, Long Beach, California, USA. PMLR.
- Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.gem-1.16>
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-3348>
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (ConvAI2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing. [https://doi.org/10.1007/978-3-030-29135-8\\_7](https://doi.org/10.1007/978-3-030-29135-8_7)

- Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? The inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020*, pages 4506–4520. International Committee on Computational Linguistics. <https://doi.org/10.18653/v1/2020.coling-main.398>
- 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.
- Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *CoRR*, abs/2104.14478. [https://doi.org/10.1162/tacl\\_a-00437](https://doi.org/10.1162/tacl_a-00437)
- Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *International Conference on Learning Representations*.
- Daphne Ippolito, Reno Kriz, João 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. <https://doi.org/10.18653/v1/P19-1365>
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *CoRR*, abs/1601.00372.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6322>
- Moin Nadeem, Tianxing He, Kyunghyun Cho, and James Glass. 2020. A systematic characterization of sampling algorithms for open-ended language generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 334–346, Suzhou, China. Association for Computational Linguistics.
- Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR’s WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.
- Myle Ott, Michael Auli, David Grangier, and Marc’Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.
- 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics. <https://doi.org/10.3115/1073083.1073135>
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6319>
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.704>
- Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3288–3294. AAAI Press. <https://doi.org/10.1609/aaai.v31i1.10984>
- Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, Copenhagen, Denmark. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1235>
- Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1331>
- Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1025>
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Kraemer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151. <https://doi.org/10.1016/j.csl.2020.101151>
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances*



- in *Neural Information Processing Systems*, volume 30.
- Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press. <https://doi.org/10.1609/aaai.v32i1.12340>
- Sean Welleck, Iliia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. 2020a. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5553–5568, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.448>
- Sean Welleck, Iliia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020b. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.
- 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 *Advances in Neural Information Processing Systems*, volume 31.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-demos.30>
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texus: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, pages 1097–1100, New York, NY, USA. Association for Computing Machinery.