

Typical Decoding for Natural Language Generation

Clara Meister¹ Tiago Pimentel² Gian Wiher¹ Ryan Cotterell^{1,2}

¹ETH Zürich ²University of Cambridge

clara.meister@inf.ethz.ch tp472@cam.ac.uk

gian.wiher@inf.ethz.ch ryan.cotterell@inf.ethz.ch

Abstract

Despite achieving incredibly low perplexities on myriad natural language corpora, today’s language models still often underperform when used to generate text. This dichotomy has puzzled the language generation community for the last few years. In this work, we posit that the abstraction of natural language as a communication channel (à la [Shannon, 1948](#)) can provide new insights into the behaviors of probabilistic language generators, e.g., why high-probability texts can be dull or repetitive. Humans use language as a means of communicating information, and do so in a simultaneously efficient and error-minimizing manner; they choose each word in a string with this (perhaps subconscious) goal in mind. We propose that generation from probabilistic models should mimic this behavior. Rather than always choosing words from the high-probability region of the distribution—which have a low Shannon information content—we sample from the set of words with information content close to the conditional entropy of our model, i.e., close to the *expected* information content. This decision criterion can be realized through a simple and efficient implementation, which we call **typical sampling**. Automatic and human evaluations show that, in comparison to nucleus and top- k sampling, typical sampling offers competitive performance in terms of quality while consistently reducing the number of degenerate repetitions.

is the choice of **decoding strategy**—i.e., the set of decision rules used to decode strings from the model. For many language generation tasks, decoding strategies which aim to find maximal probability strings produce text that is undesirable—e.g., generic or degenerate ([Holtzman et al., 2020](#); [See et al., 2019](#); [Eikema and Aziz, 2020](#); [Zhang et al., 2021](#); [DeLucia et al., 2021](#)). Rather, stochastic strategies, which take random samples from the model, often lead to text with better qualitative properties ([Fan et al., 2018](#); [Holtzman et al., 2020](#); [Basu et al., 2021](#)); yet these strategies still have their own host of problems (e.g., occasionally producing nonsensical text), while not entirely solving those seen in maximization-based approaches (e.g., falling into repetitive loops).

At first glance, it is unintuitive that high-probability strings are often neither desirable nor human-like text. A number of works have consequently concluded that there must be faults in the training objective or architecture of probabilistic language generators ([Welleck et al., 2020](#); [Guan et al., 2020](#); [Li et al., 2020](#), *inter alia*). Yet, this conclusion is at odds with these models’ performance in terms of other metrics; the fact that modern language models can achieve incredibly low perplexities suggests that they *do* provide good estimates of the probability distribution underlying human language. We posit that looking at language generation through an information-theoretic lens may shed light on this dichotomy.

Communication via natural language can intuitively be cast in the information-theoretic framework. Indeed, there is a long history of studying language via this means ([Shannon, 1948, 1951](#); [Hale, 2001](#); [Piantadosi et al., 2011](#); [Pimentel et al., 2020](#), *inter alia*). In this paradigm, strings are messages used to convey information. Each string has an associated probability of occurring, which directly reflects that string’s information

1 Introduction

Today’s probabilistic models have repeatedly demonstrated their prowess at modeling natural language, attaining low perplexities on corpora from many domains. Yet when used to generate text, their performance is far from perfect. One of the largest determinants of the generated text’s quality

content.¹ Assuming that humans use language in order to transmit information in an efficient manner (Zaslavsky et al., 2018; Gibson et al., 2019), the subset of strings typically used by humans should encode information at some (perhaps optimal) rate.² It follows that, if we want text generated from a model to be “human-like,” it should likewise adhere to this criterion. Note that high probability—or equivalently, low information—strings likely do not fall into this subset: their information content is lower than that of a typical string.

Concretely, we hypothesize that, for text to be perceived as human-like, each word should have information content close to its *expected* information content given prior context. When decoding from probabilistic language generators, we should aim to mimic this property. It turns out this is quite easy to enforce in practice: at each decoding step, we sample solely from the set of words whose negative log-probabilities are close to the conditional entropy—or equivalently, the expected information content of the subsequent word—according to our model, an operation that can be done in the same runtime as nucleus or top- k sampling. We call this new decoding strategy **typical sampling**, naming it after its relationship to the information-theoretic concept of typicality, which provides (informal) intuition for its efficacy. In experiments on summarization and story generation, we observe that, compared to nucleus and top- k sampling: (1) typical sampling reduces the number of degenerate repetitions, giving a REP value (Welleck et al., 2020) on par with human text, and (2) text generated using typical sampling is closer in quality to that of human text.³

2 Background

2.1 Probabilistic Language Generators

Systems for natural language generation are predominantly parameterized by locally-normalized probabilistic models, i.e., probability distributions q over natural language strings $\mathbf{y} = \langle y_1, y_2, \dots \rangle$

¹The Shannon information content of a message is formally defined as its negative log-probability.

²Prior works studying the uniform information density hypothesis (Levy and Jaeger, 2007; Mahowald et al., 2013; Meister et al., 2020b) observed precisely this property in humans’ use of natural language.

³An implementation of typical sampling can be found in the [Hugging Face](https://github.com/cimeister/typical-sampling.git) library. Code for reproducing experiments will be made available at <https://github.com/cimeister/typical-sampling.git>.

that are decomposed over words y_t .^{4,5}

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

The support of q is the exponentially-sized set \mathcal{Y} , which consists of all possible strings (book-ended by special beginning- and end-of-string tokens BOS and EOS) that can be constructed from words in the model’s vocabulary \mathcal{V} . In practice, we limit the set of strings we consider to $\mathcal{Y}_T \subset \mathcal{Y}$, i.e., all strings in \mathcal{Y} of some maximum length T .

The standard method for estimating the parameters of q is via maximization of the log-likelihood of a training corpus \mathcal{C} . This is equivalent to minimizing the loss

$$L(\theta; \mathcal{C}) = - \sum_{\mathbf{y} \in \mathcal{C}} \log q(\mathbf{y}) \quad (2)$$

where θ are the model parameters. It is well known that the model whose parameters minimize Equation (2) is likewise the model that minimizes the Kullback–Leibler divergence with the empirical distribution p (as defined by \mathcal{C}). Notably, this implies that the learned distribution is also optimal in an information-theoretic sense: if p represents the distribution underlying a communication channel, then minimizing $D_{\text{KL}}(p \parallel q)$ optimizes for a distribution q that places probability mass over the same messages as p .

2.2 Decoding Natural Language Strings

In short, decoding is the process of generating natural language strings from a model. While decoding is done on a word-by-word basis for all models we consider, there are still many different sets of decision rules that can be used. Given the probabilistic nature of q , a natural option would be to choose words which maximize the probability assigned by q to the resulting string. We refer to this class of methods as “mode-seeking,” as they try to find the mode of q .⁶

⁴We may also decompose strings over other units, e.g., sub-words or characters. All subsequent analyses in this work can be applied in terms of these other units.

⁵While q may be conditioned on some input \mathbf{x} , as is the case for tasks like abstractive summarization, we omit this explicit dependence for notational brevity.

⁶This is typically done in a heuristic fashion, e.g., greedily choosing the maximum probability item from $q(\cdot \mid \mathbf{y}_{<t})$ at each time step, since exactly solving for the highest probability string according to q is an NP-hard problem (Chen et al., 2018).

Yet recent research has shown that solutions to mode-seeking methods—such as greedy or beam search—are often not high-quality, even in state-of-the-art language generation models. For example, in the domain of machine translation, the most probable string under the model is often the empty string (Stahlberg and Byrne, 2019; Eikema and Aziz, 2020). For open-ended generation, mode-seeking methods produce dull, generic or even degenerate text (Fan et al., 2018; Holtzman et al., 2020).

Consequently, stochastic decoding strategies have become the mainstay for many language generation tasks. In these strategies, words are sampled randomly at each time step. While stochasticity may solve the issue of “dull or generic” text, directly sampling from $q(\cdot \mid \mathbf{y}_{<t})$ can lead to text that is incoherent and sometimes unrelated to the subject. Several works blame this behavior on attributes of the model, such as the “unreliable tail” of the distribution, perhaps caused by the non-sparse nature of the softmax transformation used to produce a probability distribution in the final layer of neural networks. They propose to fix this issue by limiting the sampling space to a core subset of words. As concrete examples, Fan et al. (2018) propose limiting the sampling space to the top- k most likely words in each decoding step; Holtzman et al. (2020) consider the smallest nucleus, i.e., subset, of words whose cumulative probability mass exceeds a chosen threshold n .

We can formalize these approaches (respectively) as alterations of $q(\cdot \mid \mathbf{y}_{<t})$ at each decoding step. Let $Z_t = \sum_{y \in \mathcal{V}^{(k)}} q(y \mid \mathbf{y}_{<t})$ where $\mathcal{V}^{(k)} \subseteq \mathcal{V}$ denotes the set of the k most likely words. **Top- k sampling** employs the truncated distribution:

$$\pi(y \mid \mathbf{y}_{<t}) = \begin{cases} q(y \mid \mathbf{y}_{<t})/Z_t, & \text{if } y \in \mathcal{V}^{(k)} \\ 0, & \text{else} \end{cases} \quad (3)$$

In a similar fashion, let $\mathcal{V}^{(n)} \subseteq \mathcal{V}$ be the smallest set such that

$$\sum_{y \in \mathcal{V}^{(n)}} q(y \mid \mathbf{y}_{<t}) \geq n \quad (4)$$

The truncated distribution for **nucleus sampling** is then computed similarly to Equation (3), albeit with $\mathcal{V}^{(n)}$ and $Z_t = \sum_{y \in \mathcal{V}^{(n)}} q(y \mid \mathbf{y}_{<t})$.

While strings generated using such stochastic methods may have lower probability according to

q , they often outperform those decoded using mode-seeking methods in terms of qualitative metrics. A number of recent works have tried to offer explanations for this phenomenon: some have attributed it to a diversity-quality trade-off (Zhang et al., 2021; Basu et al., 2021) while others blame shortcomings of model architectures or training strategies (Welleck et al., 2020; Li et al., 2020). In this work, we offer an alternative explanation, motivated by information-theory.

3 An Information-Theoretic View of Natural Language

Language is (arguably) the primary means for human communication. As such, information theory—the formal, mathematical study of communication—provides an intuitive lens through which we can study natural language. Over the past century, many insights into the development and use of language have been made using an information-theoretic framework (Hale, 2001; Aylett and Turk, 2004; Collins, 2014; Piantadosi et al., 2011; Levy, 2018; Gibson et al., 2019, *inter alia*). Building on these works, we demonstrate how concepts from information theory can help us understand certain behaviors of probabilistic language generators, and in turn, how we can generate more human-like language from them.

3.1 The Human Communication Channel

The process of communicating via natural language (either spoken or written) can be interpreted as the transmission of a message via a communication channel. In this light, a human language represents a code by which information is transmitted; a natural language string \mathbf{y} is a means of communicating some information, which we denote $I(\mathbf{y})$, and each word y_t is a symbol via which we construct our message.

Formally, information theory tells us that \mathbf{y} ’s information content can be quantified as its negative log-probability: $I(\mathbf{y}) := -\log p(\mathbf{y})$, where the distribution p is a fixed, inherent property of the communication channel. Further, if \mathbf{y} can be broken down into t units, we can express $I(\mathbf{y})$ as the sum of the information conveyed by each unit:

$$I(\mathbf{y}) = \sum_{t=1}^{|\mathbf{y}|} I(y_t) \quad (5)$$

where similarly $I(y_t) = -\log p(y_t \mid \mathbf{y}_{<t})$, the negative log-probability of y_t in its context. Note that

Equation (5) also follows from a simple application of the chain rule of probability. While we do not *a priori* know these probabilities, conveniently, they are exactly what our model q has learned to approximate, as discussed in Section 2.1.

As is typical in communication, the goal of an agent is to transmit information efficiently while also minimizing the risk of miscommunication. These goals determine how we choose to encode the information we wish to transmit,⁷ a choice encompassing which words we use and whether or not to adhere to certain paradigms that we may view as requirements, such as grammaticality. Implicit in this decision is also the amount of information transmitted by each word. When the stochastic process generating messages is stationary, we can define its formal information rate: the average amount of information transmitted by *any* symbol given all previously-generated symbols. However, language generation is arguably not a stationary process, i.e., we cannot say there is necessarily some t for which all words $< t$ do not affect the distribution over y_t . We can nonetheless still compute the expected amount of information a *specific* symbol y_t in our message will contain:

$$\mathbb{E}[I(y_t)] = -\sum_{y \in \mathcal{V}} p(y | \mathbf{y}_{<t}) \log p(y | \mathbf{y}_{<t}) \quad (6)$$

Note that this quantity is simply the conditional entropy of $p(\cdot | \mathbf{y}_{<t})$, which we denote as $H(p(\cdot | \mathbf{y}_{<t}))$.⁸

So how do we, as humans, choose the amount of information we transmit over the course of a natural language string? This is a research question in and of its own right, one without a concrete answer yet. Research in psycholinguistics, however, suggests that speakers avoid producing words with either very high or very low information content (Levy and Jaeger, 2007; Frank and Jaeger, 2008; Mahowald et al., 2013). Furthermore, cross-linguistic research has shown that languages trade-off information content and speech rate, perhaps aiming at a specific information rate value (Coupé et al., 2019; Pimentel et al., 2021). It seems, thus, that a core component of what makes text

⁷See Gibson et al. 2019 for an in-depth review of how efficiency has shaped the evolution of language.

⁸We use the convention of supplying the entropy function H with a probability distribution—rather than a random variable—to make explicit the distribution we are operating over.

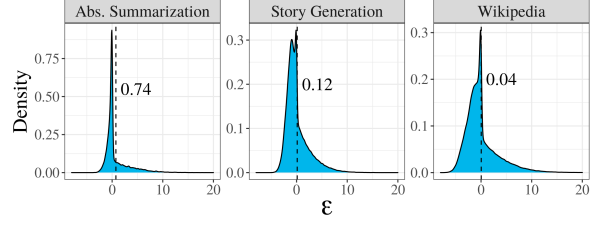


Figure 1: The per-token distribution of the deviation (ε) of information content from conditional entropy. Values are computed using the reference (human) text for three different language generation tasks, where probabilities and entropies are computed using probabilistic models trained on the respective task (see Section 5 for model details). Dashed line and label indicate mean ε . Per token distributions of conditional entropies and information contents are shown in Appendix B for reference.

“human-like” is its per-word information content. Consequently, in this work, we posit the following:

Hypothesis 3.1. *Any given word should have an information content close to the expected information content, i.e., the conditional entropy given prior context. In other words, we expect the difference:*

$$\varepsilon = |H(p(\cdot | \mathbf{y}_{<t})) - I(y_t)| \quad (7)$$

to be small in human-like text for all words y_t .

We provide empirical motivation for our hypothesis in Figure 1, which shows—for human-generated text—the distribution of the deviation of $I(y_t)$ from $H(p(\cdot | \mathbf{y}_{<t}))$ under a model q which serves as an estimate of p . There are two important observations that can be taken from this figure: (1) the peaked nature of the distributions reveals that humans indeed tend to form language with per-word information content quite close to the *expected* information content and (2) the centering of these distributions around a value close to 0 reveals that our probabilistic language generators are learning what this rate is.

3.2 Relationship of Hypothesis 3.1 to Typicality

Hypothesis 3.1 can intuitively be linked to the notion of **typicality** in information theory (Shannon, 1948). Typicality is a property of messages from a specific stochastic process: typical messages are the ones that we would expect from this process, given its probability distribution. These sets of messages have a quantifiable attribute—their average per-symbol information content is close to the entropy rate of their source. Formally, an ε -typical

message of length N from a stochastic process defined by the distribution p has probability within the range

$$2^{-N(H(p)+\varepsilon)} \leq p(y_1, \dots, y_N) \leq 2^{-N(H(p)-\varepsilon)} \quad (8)$$

for a given ε . Interestingly, this definition implies that the highest probability message (often) would not be considered “typical”—its average information content is too low.

While the formal notion of typicality is not directly applicable in this context,⁹ it is illustrative of why Hypothesis 3.1 might indeed be true: it demonstrates a concrete relationship between messages we expect to see—given the distribution underlying a stochastic process—and the information content of those messages. Analogously, our hypothesis predicts what information content we should expect in messages produced by the distribution over human language. However, typicality is a property with respect to the information content of a message in its entirety, not describing characteristics of individual symbols in the message. Hypothesis 3.1 is instead more akin to a “local” typicality, where individual symbols themselves have information content within a specified range. The motivation for this more local definition is that natural-sounding language should be typical everywhere; it should not be able to compensate for unusually low probability in the first half, e.g., grammatical errors, with unusually high probability in the second half, e.g., especially frequent words.

This local definition also has a more direct link to several psycholinguistic theories, which we discuss next. Further, as we will see in Section 4, it proves more practically useful when creating decision rules for decoding from standard probabilistic models, which must be performed word-by-word.

3.3 Relationship of Hypothesis 3.1 to Psycholinguistic Concepts

We next motivate this hypothesis as an extension of two psycholinguistic theories, namely: the uniform information density hypothesis, and the rational speech act.

⁹Typicality is defined in terms of stationary ergodic processes. Generation from standard neural language models can not be characterized as such a process due to the absorbing nature of the EOS state and the ability of recurrent neural networks to encode arbitrarily long sequences.

The Uniform Information Density Hypothesis. The uniform information density (UID) hypothesis (Fenk and Fenk, 1980; Levy and Jaeger, 2007) states that speakers construct their utterances such that information is distributed uniformly across them. While this hypothesis may allow multiple interpretations (see, e.g., Meister et al., 2021, for discussion), a predominant one is that speakers optimize their sentences to maximize the use of a communication channel—choosing words such that their information rate is closer to a target channel capacity. If a word conveys more information than the channel allows, there is a risk for miscommunication; simultaneously, if a word conveys very low information, then this channel is being used inefficiently. Analogously, we propose here that speaker’s will avoid producing words with “out of the ordinary” information content, thus choosing words close to the expected information at each moment.

A UID-inspired objective has previously been shown to improve language modeling results (Wei et al., 2021). Further, the UID hypothesis has been used as rationale for the effectiveness of beam search in the context of machine translation (Meister et al., 2020b). While at first, the hypothesis presented in this work may seem at odds with results showing the efficacy of mode-seeking decoding strategies, like beam search, a closer look reveals they are in fact compatible. When trained *without* label-smoothing, which artificially inflates conditional entropies, machine translation models tend to have quite low conditional entropies (see e.g., Fig. 3 in Meister et al., 2020a). Therefore, at each decoding step, the set of words with negative log-probability near the conditional entropy of the model are typically those with high probability—the same as those chosen by beam search.

The Rational Speech Act. The rational speech act (RSA; Frank and Goodman, 2012) is a recently-proposed framework which models a speaker’s pragmatic behavior. In short, RSA casts a speaker’s behavior as the maximization of a utility function: a string’s usefulness to its listener. More formally, RSA introduces the concept of a literal speaker, who produces strings y according to a base (naïve) distribution p_0 . The value a listener can extract from a message is then a function of that entire distribution $u(y; p_0)$, i.e., the listener’s utility function. Finally, the pragmatic speaker produces strings to maximize this utility, as

opposed to following its expected literal behavior. Within this framework, we posit that a listener’s utility is modulated by how close to the entropy each word’s information content is, i.e.:

$$u(y_t; p) \propto -|H(p(\cdot | \mathbf{y}_{<t})) - I(y_t)| \quad (9)$$

and that pragmatic speakers will only produce strings where the utility is larger than a specific value at each time step. In other words, a speaker will produce strings whose per-word information contents are close to what is expected by the listener.

4 An Information-Theoretic Decoding Strategy

Considering probabilistic language generators in the information-theoretic framework leads to a number of insights into behaviors that have been previously observed when decoding text from these models. In conjunction with the rationale and observations in Section 3, it also motivates a new decoding strategy, which we call **typical sampling**.

4.1 Understanding Probabilistic Language Generators

A probabilistic language generator q approximates the empirical distribution p .¹⁰ Let us now adopt the interpretation that this distribution underlies a natural language communication channel. This framing offers an informal explanation between some qualitative properties of strings and their probability under a model q , which we believe may reconcile why models performing so well (in terms of metrics such as perplexity) can still exhibit such “undesirable” behavior when used to generate text.

First, the connection between probability and information content may explain why high-probability text is often dull or generic (Holtzman et al., 2020; Eikema and Aziz, 2020)—its low information content likely makes for boring or uninformative text. This connection also offers a potential explanation for the rather strange behavior that, when generations fall into a repetitive loop, language models often assign increasingly higher probability to the repeated substring (Holtzman et al., 2020)—the substring conveys less and less information after each occurrence.

¹⁰Our following analysis assumes q is a perfect representation of p , even though it is undoubtedly not. Still, these models provide an incredibly good approximation of p —as shown by performance w.r.t. metrics such as perplexity.

A further implication of this framing is the equivalence between decoding strings from a probabilistic language generator q and sampling messages from the natural language communication channel. If we wish to solely sample from the subset of messages that a human would typically construct, i.e., that are “human-like,” then we should begin by narrowing down this subset to those messages that meet at least some of the same criteria as human-generated messages. In this work, one such criterion we have identified is that per-word information content lies close to the expected information content of the symbol: $I(y_t) \approx \mathbb{E}[-\log p(\cdot | \mathbf{y}_{<t})]$.

Notably, we already see motivation for this criterion in the performance of several well-known decoding strategies. For example, beam search is the predominant decoding strategy for machine translation models (Wu et al., 2016; Edunov et al., 2018; Ng et al., 2019), a setting in which it (incidentally) often already enforces this criterion (see Section 3.3 for elaboration). Yet when used in more open-ended tasks, where models typically have higher entropies, beam search can lead to low-quality text (Li et al., 2016; Holtzman et al., 2020; Welleck et al., 2020). As another example, Mirostat (Basu et al., 2021) decodes strings with a target perplexity, or equivalently, an overall target per-word information content. Their approach alleviates some of the short-comings of top- k and nucleus sampling. These observations motivate a new decoding strategy in which our information-theoretic criterion is explicitly enforced, which we subsequently present.

4.2 Typical Sampling

We define $\mathcal{V}^{(\tau)} \subseteq \mathcal{V}$ as the subset of words that minimize

$$\sum_{y \in \mathcal{V}^{(\tau)}} |H(q(\cdot | \mathbf{y}_{<t})) + \log q(y | \mathbf{y}_{<t})| \quad (10)$$

s.t. $\sum_{y \in \mathcal{V}^{(\tau)}} q(y | \mathbf{y}_{<t}) \geq \tau$. That is, we limit our sampling distribution to only those words with negative log-probability within a certain absolute range from the conditional entropy of the model at that time step. In the spirit of nucleus sampling, this range is determined by a hyperparameter τ , the amount of probability mass from the original distribution that we wish to consider. We then renormalize our truncated distribution as in Equation (3), where similarly we have $Z_t = \sum_{y \in \mathcal{V}^{(\tau)}} q(y | \mathbf{y}_{<t})$.

Note that this rule does *not* imply that high-probability words should not be chosen. Indeed,

	Story Generation						
	PPL (g)	PPL (i)	MAUVE (\uparrow)	REP (\downarrow)	Zipf	D (\uparrow)	Human (\uparrow)
Reference	16.33	3.19	—	0.28	1.09	0.85	4.12 (± 0.02)
Temperature ($\tau=0.5$)	25.34 ($+9.01$)	2.90 (-0.29)	0.95	0.25	1.07 (-0.02)	0.87	4.13 (± 0.02)
Temperature ($\tau=1$)	25.67 ($+9.34$)	2.39 (-0.8)	0.95	0.26	1.07 (-0.02)	0.87	4.13 (± 0.02)
Nucleus ($n=0.9$)	7.75 (-8.58)	2.23 (-0.96)	0.95	0.35	1.29 ($+0.20$)	0.79	4.09 (± 0.02)
Nucleus ($n=0.95$)	11.65 (-4.68)	2.39 (-0.80)	0.95	0.30	1.20 ($+0.11$)	0.84	4.13 (± 0.02)
Top- k ($k=30$)	7.07 (-9.26)	2.90 (-0.29)	0.88	0.35	1.41 ($+0.32$)	0.80	4.13 (± 0.02)
Top- k ($k=40$)	11.83 (-4.5)	2.50 (-0.69)	0.92	0.35	1.33 ($+0.24$)	0.82	4.09 (± 0.02)
Mirostat ($\tau=3$)	8.14 (-8.19)	3.13 (-0.06)	0.93	0.34	1.30 ($+0.21$)	0.83	4.12 (± 0.02)
Typical ($\tau=0.2$)	14.25 (-2.08)	3.13 (-0.06)	0.78	0.30	1.27 ($+0.18$)	0.84	4.15 (± 0.02)
Typical ($\tau=0.95$)	11.59 (-4.74)	2.39 (-0.80)	0.96	0.31	1.21 ($+0.12$)	0.84	4.13 (± 0.02)

Table 1: Automatic quality and diversity metrics, as described in Section 5.1, along with human ratings on the WRITINGPROMPTS dataset. Human ratings are averaged across criteria to form a single metric. Bolded values are the best results among decoding strategies, where for perplexity and Zipf’s coefficient, we take this to be the delta from measurements on human text (numbers in purple). Numbers in blue are standard error estimates. Results are from fine-tuned GPT-2 large; results for GPT-2 medium are given in Appendix B.

in the situation where conditional entropy is low—i.e., when the model places most of the probability mass on a small subset of words—then it is likely the case that only high-probability words y_t meet the criterion $I(y_t) \approx H(q(\cdot | \mathbf{y}_{<t}))$.

Computational Complexity. From a practical perspective, this scheme can be implemented with the same efficiency as nucleus or top- k sampling. First, we compute the conditional entropy, which is an $\mathcal{O}(|\mathcal{V}|)$ operation. Second, we sort words by their absolute distance from $H(q(\cdot | \mathbf{y}_{<t}))$, which can be done in $\mathcal{O}(|\mathcal{V}| \log |\mathcal{V}|)$ time with standard sorting algorithms. Third, similarly to nucleus sampling, we greedily take words from this list until their cumulative probability exceeds the threshold τ , which again takes $\mathcal{O}(|\mathcal{V}|)$ time. Thus, creating our altered distribution has time complexity $\mathcal{O}(|\mathcal{V}| \log |\mathcal{V}|)$.

5 Experiments

In this section, we explore the efficacy of our decoding strategy for two natural language generation tasks: abstractive summarization and story generation. We assess performance with respect to several other stochastic decoding strategies: nucleus sampling, top- k sampling, temperature sampling, beam search and Mirostat. Our evaluation includes both automatic metrics and human ratings.

5.1 Setup

Models and Data. We use the [Hugging Face](#) framework ([Wolf et al., 2020](#)) for reproducibility, employing their implementations of nucleus, top- k

and temperature sampling and beam search. We rely on the implementation of Mirostat provided by the authors.¹¹ For story generation, we fine-tune the medium and large versions of GPT-2 ([Radford et al. 2019](#); from checkpoints made available by OpenAI) on the WRITINGPROMPTS dataset ([Fan et al., 2018](#)). We use the medium checkpoint, albeit fine-tuned on WIKITEXT-103 ([Merity et al., 2017](#)) to produce the data used in Figure 1. For abstractive summarization, we use BART ([Lewis et al., 2020](#)) fine-tuned on the CNN/DAILYMAIL dataset ([Nallapati et al., 2016](#)).¹² All reported metrics are computed on the respective test sets.

Hyperparameters In a preliminary hyperparameter sweep using MAUVE¹³ ([Pillutla et al., 2021](#)), we found $k = 30, 40$, $n = 0.9, 0.95$ and $\tau = 3.0$ to perform best for top- k sampling, nucleus sampling and Mirostat, respectively. For typical sampling, we found $\tau = 0.2, \tau = 0.95$ to provide the best results for story generation and abstractive summarization, respectively. Standard values according to the literature for other hyperparameters (i.e., for beam search and temperature sampling) were employed. We use these hyperparameter values in our human evaluations and in computation of automatic metrics.

¹¹<https://github.com/basusourya/mirostat>

¹²As we are interested in getting as close an estimate of p as possible with our models q , all fine-tuning is done *without* label-smoothing.

¹³We use the default settings given by the authors for all MAUVE computations, albeit we employ different LMs in our parameter sweep vs. reported results (standard GPT-2 vs. GPT-2 large, respectively) to reduce bias in the final results.

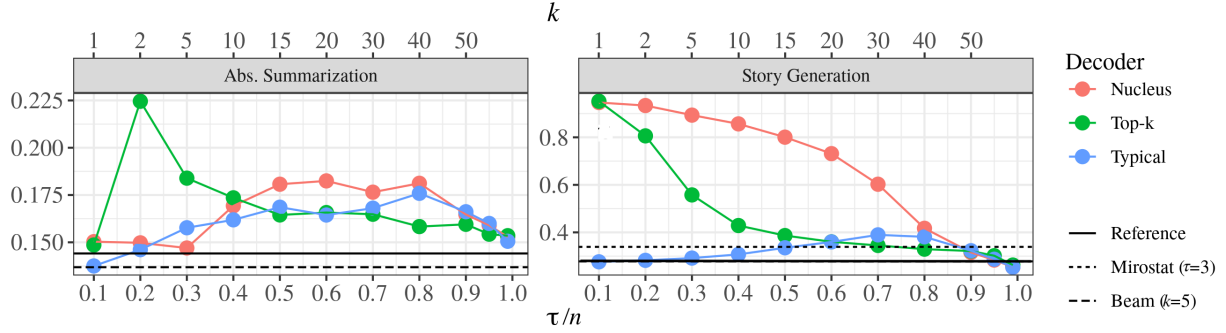


Figure 2: REP (Welleck et al., 2020) values for different k and τ/n (lower is better). Lines indicate REP measurement for reference text and Mirostat/beam search.

Abstractive Summarization							
	PPL (g)	PPL (i)	MAUVE (\uparrow)	REP (\downarrow)	Zipf	D (\uparrow)	Human (\uparrow)
Reference	10.29	3.35	—	0.13	0.76	0.97	4.31 (± 0.03)
Beam ($k=5$)	1.39 (-8.90)	3.35 ($+0.00$)	0.90	0.14	0.77 ($+0.01$)	0.97	4.35 (± 0.03)
Temperature ($\tau=0.5$)	7.10 (-3.19)	3.73 ($+0.38$)	0.97	0.15	0.75 (-0.01)	0.97	4.25 (± 0.03)
Temperature ($\tau=1$)	6.46 (-3.83)	3.35 ($+0.00$)	0.95	0.14	0.75 (-0.01)	0.97	4.29 (± 0.03)
Nucleus ($n=0.9$)	2.97 (-7.32)	3.27 (-0.08)	0.90	0.17	0.93 ($+0.17$)	0.96	4.26 (± 0.03)
Nucleus ($n=0.95$)	3.96 (-6.33)	3.64 ($+0.29$)	0.99	0.15	0.91 ($+0.15$)	0.97	4.26 (± 0.03)
Top- k ($k=30$)	3.13 (-7.16)	3.27 (-0.08)	0.98	0.16	0.93 ($+0.17$)	0.97	4.31 (± 0.03)
Top- k ($k=40$)	3.26 (-7.03)	3.12 (-0.23)	0.96	0.16	0.93 ($+0.17$)	0.97	4.29 (± 0.03)
Typical ($\tau=0.2$)	3.80 (-6.49)	3.79 ($+0.44$)	0.72	0.14	0.91 ($+0.15$)	0.97	4.27 (± 0.03)
Typical ($\tau=0.95$)	3.86 (-6.43)	3.64 ($+0.29$)	0.96	0.15	0.92 ($+0.16$)	0.97	4.32 (± 0.03)

Table 2: Automatic quality and diversity metrics, as described in Section 5.1, along with human ratings on the CNN/DAILYMAIL dataset. Human ratings are averaged across criteria to form a single metric. Bolded values are the best results among decoding strategies, where for perplexity and Zipf’s coefficient, we take this to be the delta from measurements on human text (numbers in purple). Numbers in blue are standard error estimates.

Automatic Quality Metrics. As automatic quality metrics, we evaluate the generated text’s perplexity—under both the model used to generate the text ($\text{PPL}(g)$) and an independent LM ($\text{PPL}(i)$). Several prior works have shown that neither low nor high perplexity (Zhang et al., 2021; Nadeem et al., 2020; Pillutla et al., 2021) are direct indicators of text quality. Rather, human-like text often has perplexity within a certain range. Consequently, we report the difference in this metric from the reference text as well. We additionally compute MAUVE scores (Pillutla et al., 2021) with the reference text—a metric that uses LM embeddings to measure the similarity between two text distributions. For computing MAUVE, we use the implementation provided by the authors.

Automatic Diversity Metrics. For automatic diversity metrics, we compute REP (Welleck et al., 2020), Zipf’s coefficient, and n -gram diversity. For REP we use the average of REP/ℓ scores (as defined in eq. 9 of Welleck et al., 2020) for

$\ell \in \{16, 32, 128\}$. We define n -gram diversity D as the average fraction of unique vs. total n -grams for $n = 1, 2, 3, 4$ in a string:

$$D = \sum_{n=1}^4 \frac{\# \text{unique } n\text{-grams in string}}{\# n\text{-grams in string}} \quad (11)$$

Human Evaluations. We use Amazon Mechanical Turk to obtain human judgments of text quality from 5 different annotators on 200 examples per decoding strategy—per task. We largely follow DeLucia et al. (2021) in setting up our evaluations: For abstractive summarization, we ask annotators to score on fluency and relevance while for story generation, annotators score on fluency, coherence, and interestingness. We choose these criteria following recommendations from van der Lee et al. (2019). We use a 5-point Likert scale for each criterion, giving detailed examples and descriptions for each of the options. More details on setup can be found in Appendix A, where we provide the exact instructions presented to the workers.

Abstractive Summarization (CNN/DailyMail)

Prompt	(CNN) The attorney for a suburban New York cardiologist charged in what authorities say was a failed scheme to have another physician hurt or killed is calling the allegations against his client “completely unsubstantiated.” Appearing Saturday morning on CNN’s “New Day,” Randy Zelin defended his client, Dr. Anthony Moschetto, who faces criminal solicitation, conspiracy, burglary, arson, criminal prescription sale and weapons charges in connection to what prosecutors called a plot to take out a rival doctor on Long Island. “None of anything in this case has any evidentiary value,” Zelin told CNN’s Christi Paul...
Reference	A lawyer for Dr. Anthony Moschetto says the charges against him are baseless. Moschetto, 54, was arrested for selling drugs and weapons, prosecutors say. Authorities allege Moschetto hired accomplices to burn down the practice of former associate.
Beam $k = 5$	Dr. Anthony Moschetto faces criminal solicitation, conspiracy, burglary, arson and weapons charges. “None of anything in this case has any evidentiary value,” his attorney says.
Nucleus $n = 0.95$	Dr. Anthony Moschetto, 54, pleaded not guilty to charges Wednesday. Two men – identified as James Chmela and James Kalamaras – were named as accomplices.
Top-k $k = 30$	Dr. Anthony Moschetto is accused of providing police with weapons and prescription drugs. Authorities say he was part of a conspiracy to harm or kill a rival doctor. His attorney calls the allegations against his client “completely unsubstantiated”
Typical $\tau = 0.95$	Dr. Anthony Moschetto is charged with crimes including arson, conspiracy, burglary, prescription sale, weapons charges. His attorney says “none of anything in this case has any evidentiary value”

Table 3: Sample generations for abstractive summarization; examples correspond to ID 1 in the test set. Decoding strategy hyperparameters are chosen based off of performance in human evaluations shown in Table 2.

5.2 Results

Quantitative Performance. Tables 1 and 2 show the results of our different evaluation metrics. Human scores are averaged across the qualitative metrics to give an aggregate score; the value in parentheses is the standard error of the estimate. We show full breakdowns of score distributions in Table 5. We see that in general, typical sampling performs on par with or better than other sampling techniques, producing text with human quality ratings closest to that of the reference among the stochastic decoding strategies. Figure 2 shows REP measurements for different values of the hyperparameters k , n , and τ for top- k , nucleus, and typical sampling, respectively. Interestingly, REP appears to be far less sensitive to τ than to k and n . While many values of k and n appear to lead to degenerate repetitions in story generation, most values of τ lead to text with a REP value on par with human text.

Qualitative Performance. We provide examples of text generated according to each of the decoding strategies in Tables 3 and 4. For text that is too long, we provide the full version in Table 7 in the appendix. For both of the tasks, we choose the example with ID 1 in the respective test set and pro-

vide examples from each of the decoding strategies, employing the hyperparameter values that lead to the best human scores in Tables 1 and 2, where ties are broken based on performance in other criteria.

For the summarization task, we see that typical sampling provides a comprehensive and coherent summary of the article, quite similar to that of beam search. In comparison, the text produced by temperature sampling is not necessarily coherent; text from nucleus sampling and top- k sampling misses some of the important information in the article, e.g., the charges of burglary and arson.

While the qualitative performance of story generation are much more subjective, typical sampling arguably provides the most fluent story amongst all the decoding strategies. Other stories lack coherency and even within the first few sentences, we see repeated phrases and words. Together, these results suggest that typical sampling may indeed produce more desirable text.

6 Conclusion

In this work, we analyze decoding from probabilistic language generators using an information-theoretic framework. Under this paradigm, a language model approximates the probability distribution that defines a communication channel; gener-

Story Generation (WritingPrompts)

Prompt	A kid doodling in a math class accidentally creates the world’s first functional magic circle in centuries.
Reference	It was dark and Levi was pretty sure he was lying on his back. There was firelight flickering off of what was left of a ceiling. He could hear something but it was muffled. He was almost positive it was screaming. When he tried to move he felt an excruciating pain in his left side that caused him to cry out...
Nucleus $n = 0.95$	“Why do you draw a circle?” He asked as he handed the piece of paper over to her. She rolled her eyes and looked down at the picture, “Because I like math!” “Then what?” “I like to draw the circles.” She ran a hand through her curls and looked away. “See, I have three circles.” “Hmph, I didn’t think I saw that in your class.” A younger student brought up the ceiling and murmured...
Top-k $k = 30$	“Aha. Aah! That makes three.” Professor Gosh sighed. “Oh my God, you had to have been there.” “Well, I’m not sure if that’s the right term, but this circle is going to work.” “But I just told you that it wasn’t working, didn’t you, you little sh**?” “I’m not kidding! That circle is gon na work!” “So what now?”...
Temp $\tau = 1.0$	It was teachers duty to ensure that our kingdom is pure and pure and successful, however it does not mean we do not try to be as rebellious as any other kingdom. Most of our teachers are suitable for rulership, being knowledgeable in sadistic rec thinking of time stopping magic circle created by several apprentices...
Mirostat $\tau = 3$	ready to go home, when a little kid starts to sketch a circle. “No. Just stop. It’s not working.” “But the circle turns to smoke and then to a flame. It’s not working, I know it is!” Another kid, who was also drawing a circle with his phone, starts to scream. The teacher, looking at the circle, lost his patience and started to yell...
Typical $\tau = 0.2$	As soon as he got the clock on his desk he jumped up. The noise it was making had changed his day forever. His hair, normally unkempt, now glistened and he had to have more hair because of it. “Yes!” He had declared triumphantly as he watched his father fumble around in the backpack in search of a little compass he always kept with him. He took it from the desk and quickly walked out the door to see what had happened. He wasn’t the first child in the world to draw something...

Table 4: Sample generations for story generation from GPT-2 large fine-tuned on the WRITINGPROMPTS dataset; examples correspond to ID 1 in the test set. Decoding strategy hyperparameters are chosen based off of performance in human evaluations shown in Table 1. Full stories shown in Appendix B.

ating text from such a model is then equivalent to sampling messages from this channel. Motivated by results in psycholinguistics and pragmatics, we hypothesize that—with the goal of communicating efficiently—humans produce text whose per-word information content is within a close range of the *expected* information content of a word given prior context. This observation provides a simple new criterion for decoding more human-like text from probabilistic language generators. In experiments on two language generation tasks, we find that our strategy—called typical sampling—leads to text of comparable or better quality than other stochastic decoding strategies according to human ratings. Further, when compared to these other decoding strategies, several quantitative properties of typically sampled text more closely align with those of human text.

References

- Matthew Aylett and Alice Turk. 2004. [The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech](#). *Language and Speech*, 47(1):31–56.
- Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. [Mirostat: A perplexity-controlled neural text decoding algorithm](#). In *9th International Conference on Learning Representations*.
- Yining Chen, SORCHA Gilroy, Andreas Maletti, Jonathan May, and Kevin Knight. 2018. [Recurrent neural networks as weighted language rec-](#)

- ognizers. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2261–2271, New Orleans, Louisiana. Association for Computational Linguistics.
- Michael Xavier Collins. 2014. [Information density and dependency length as complementary cognitive models](#). *Journal of Psycholinguistic Research*, 43(5):651–681.
- Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. [Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche](#). *Science Advances*, 5(9).
- 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.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- 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*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- 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.
- August Fenk and Gertraud Fenk. 1980. Konstanz im kurzzeitgedächtnis-konstanz im sprachlichen informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.
- A. F. Frank and T. Jaeger. 2008. [Speaking rationally: Uniform information density as an optimal strategy for language production](#). In *the 30th Annual Meeting of the Cognitive Science Society*.
- Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.
- Edward Gibson, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. [How efficiency shapes human language](#). *Trends in Cognitive Sciences*.
- Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. [A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation](#). *Transactions of the Association for Computational Linguistics*, 8:93–108.
- John Hale. 2001. [A probabilistic Earley parser as a psycholinguistic model](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR*.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Roger Levy. 2018. [Communicative efficiency, Uniform Information Density, and the Rational Speech Act Theory](#). In *the 40th Annual Meeting of the Cognitive Science Society*, pages 684–689. Cognitive Science Society.
- Roger Levy and T. Florian Jaeger. 2007. [Speakers optimize information density through syntactic reduction](#). In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.
- 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.
- 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.
- Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. [Don’t say that! making inconsistent dialogue unlikely with unlikelihood training](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.
- Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. [Info/information theory: Speakers choose shorter words in predictive contexts](#). *Cognition*, 126(2):313–318.
- Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. [Revisiting the Uniform Information Density hypothesis](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020a. [Generalized entropy regularization or: There’s nothing special about label smoothing](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6870–6886, Online. Association for Computational Linguistics.
- Clara Meister, Tim Vieira, and Ryan Cotterell. 2020b. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. [Pointer sentinel mixture models](#). In *Proceedings of the 5th International Conference on Learning Representations, Conference Track Proceedings*, Toulon, France.
- 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.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. 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.
- Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. [Word lengths are optimized for efficient communication](#). *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. [MAUVE: Measuring the gap between neural text and human text using divergence frontiers](#). In *Advances in Neural Information Processing Systems*.
- Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. [A surprisal–duration trade-off across and within the world’s languages](#). In *Proceedings of the 2021 Conference on Empirical*

- Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. [Phonotactic Complexity and Its Trade-offs](#). *Transactions of the Association for Computational Linguistics*, 8:1–18.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. [Do massively pretrained language models make better storytellers?](#) In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics.
- C. E. Shannon. 1951. [Prediction and entropy of printed english](#). *Bell System Technical Journal*, 30(1):50–64.
- Claude E. Shannon. 1948. [A mathematical theory of communication](#). *Bell System Technical Journal*, 27:623–656.
- 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.
- Jason Wei, Clara Meister, and Ryan Cotterell. 2021. [A cognitive regularizer for language modeling](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5191–5202, Online. Association for Computational Linguistics.
- Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. [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.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. [Google’s neural machine translation system: Bridging the gap between human and machine translation](#). *CoRR*, abs/1609.08144.
- Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. [Efficient compression in color naming and its evolution](#). *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.
- 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.

A Human Evaluation Setup

We use [Amazon Mechanical Turk](#) framework for collecting human ratings of text. We use solely MTurk Master Workers in order to maximize the quality of our ratings. For story generation and abstractive summarization, each Human Intelligence Task (HIT) consists of either a single prompt from which a story should be generated or a single news article to be summarized. The raters are first presented with the different rating criteria, along with descriptions of the type of text that meets these criteria at different levels of the scale. These definitions can be seen in Figures 3 and 4. Raters are additionally provided several examples of stories/summarizations meeting/failing to meet the rating criteria. They are then presented with the respective prompt/news article and the corresponding stories/summaries generated by different decoders and by the reference in random order. We use an attention check in each HIT. Responses where the attention check has been failed are thrown out. For each of the rating criteria, the rater assigns a score from 1 to 5. For each story/summarization and each of the criteria, we take the median score across raters as the final respective score. Statistics for these scores can be seen in Table 5. Workers are awarded \$1.50 per HIT for the abstractive summarization task and \$2 per HIT for the story generation task, for which entries were longer. These rates translate to >\$15/hour.

Detailed Instructions	Examples
<p>We will reject your HIT if you fail attention checks or if you have unusually low agreement with other annotators.</p> <h3>Definitions</h3> <p>Below you will find multiple prompts and stories (narratives) generated from those prompts. Please rate the stories according to their interestingness, fluency and coherence following the given definitions and examples. We will reject your HIT if you input obviously wrong answers. The 5-point scale for each definition should be used as a guideline. <u>The definitions are displayed when hovering over each radio button for convenience.</u> (Note: if the definitions do not appear even after a few seconds, please leave your browser (e.g. Chrome) and OS (e.g. Windows) information in the comment box.)</p> <ul style="list-style-type: none">• Interesting: The story is fun to read. It feels creative, original, dynamic, and/or vivid. The opposite of this might be something that's obvious, stereotypical/unoriginal, and/or boring.<ul style="list-style-type: none">◦ <u>Very interesting:</u> The story has themes, characters, and dialog that make you want to keep reading it and you might even want to show it to a friend◦ <u>Somewhat interesting:</u> The story has themes, characters, dialog, and/or a writing style that pique your interest◦ <u>Mildly interesting:</u> There are moments of interest but the story is not too notable.◦ <u>Not very interesting:</u> You finish the story but can't remember anything unique about it. Adequate, but not a fun read◦ <u>Not at all interesting:</u> You do not even want to finish reading the story. It is boring and/or unoriginal.• Fluent: The story is written in grammatical English. No obvious grammar mistakes that a person wouldn't make. An incomplete final word or incomplete sentence does not count as a mistake and should not affect fluency. The English sounds natural. Note: do not take off points for spaces between punctuation (e.g. "don't") and simpler sentences. Simple English is as good as complex English, as long as everything is grammatical.<ul style="list-style-type: none">◦ <u>Very fluent:</u> The sentences read as if they were written by a native English speaker with 1 or no errors.◦ <u>Somewhat fluent:</u> The sentences read as if they were written by a native English speaker with very few errors. Some minor mistakes that a person could have reasonably made.◦ <u>Mildly fluent:</u> The sentences could have been written by a human but it is not entirely obvious.◦ <u>Not very fluent:</u> Many sentences have frequently repeated words and phrases. Obvious mistakes.◦ <u>Not at all fluent:</u> The sentences are completely unreadable. If the same sentence is repeated over and over for the entire story, that story is considered not at all fluent.• Coherent: The story feels like one consistent story, and not a bunch of jumbled topics. Stays on-topic with a consistent plot, and doesn't feel like a series of disconnected sentences.<ul style="list-style-type: none">◦ <u>Very coherent:</u> The sentences when taken as a whole all have a clearly identifiable plot◦ <u>Somewhat coherent:</u> Many of the sentences work together for a common plot with common characters. One or two unrelated sentences.◦ <u>Mildly coherent:</u> Around half of the sentences work together. The plot is not entirely clear though.◦ <u>Not very coherent:</u> Only a few sentences seem to be from the same story; the others are random.◦ <u>Not at all coherent:</u> There is absolutely no identifiable plot. Each sentence feels completely disconnected from every other sentence. <hr/> <p>Please confirm the following worker criteria:</p> <ul style="list-style-type: none"><input type="checkbox"/> I have read the instructions<input type="checkbox"/> I have read the examples<input type="checkbox"/> I am a native English speaker	

Figure 3: Stories survey.

Detailed Instructions

Examples

We will reject your HIT if you fail attention checks or if you have unusually low agreement with other annotators.

Definitions

Below you will find a news article and multiple summaries of this article. The news article is repeated for each summary for reference. Please rate the summaries according to their fluency and relevance following the given definitions and examples. We will reject your HIT if you input obviously wrong answers. The 4-point scale for each definition should be used as a guideline. The definitions are displayed when hovering over each radio button for convenience. (Note: if the definitions do not appear even after a few seconds, please leave your browser (e.g. Chrome) and OS (e.g. Windows) information in the comment box.)

- Fluent:** The summary is written in grammatical English. No obvious grammar mistakes that a person wouldn't make. **An incomplete final word or incomplete sentence does not count as a mistake and should not affect fluency.** The English sounds natural. Note: do not take off points for spaces between punctuation (e.g. "don't") and simpler sentences. Simple English is as good as complex English, as long as everything is grammatical.
 - Very fluent: The sentences read as if they were **written by a native English speaker with 1 or no errors.**
 - Somewhat fluent: The sentences read as if they were written by a **native English speaker with very few errors.** Some minor mistakes that a person could have reasonably made.
 - Mildly fluent: The sentences could have been written by a human but it is not entirely obvious as there are a number of mistakes.
 - Not very fluent: Many sentences have **frequently repeated words and phrases.** Obvious mistakes.
 - Not at all fluent: The sentences are **completely unreadable.** If the same sentence is **repeated over and over** for the entire story, that story is considered not at all fluent.
- Relevant:** The summary captures all the relevant events, persona and locations from the article.
 - Very relevant: It is very clear the summary captures the theme, vocabulary, and specific persona and locations from the article.
 - Somewhat relevant: Most sentences include relevant points from the article.
 - Mildly relevant: A few sentences include relevant points from the article.
 - Not very relevant: The summary mentions something from the article but contains mostly unrelated text.
 - Not at all relevant: It is as if the summary was written without reading the article.

Please confirm the following worker criteria:

☐ I have read the instructions
☐ I have read the examples
☐ I am a native English speaker

Figure 4: Summarization survey.

B Additional Results

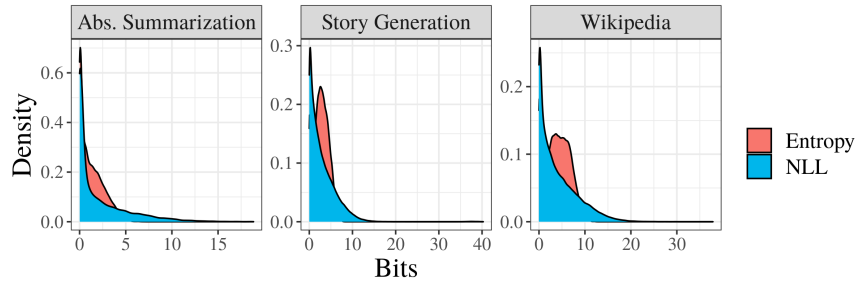


Figure 5: Distributions of conditional entropies and information contents per token for three different language generation tasks for human text, i.e., the reference text for each of the respective datasets.

Decoder	Story Generation (l)			Story Generation (m)			Summarization	
	Coherence	Fluency	Interestingness	Coherence	Fluency	Interestingness	Fluency	Relevance
Reference	4.36 (± 0.31)	4.25 (± 0.23)	4.56 (± 0.25)	4.02 (± 0.27)	4.2 (± 0.27)	4.15 (± 0.2)	4.43 (± 0.25)	4.18 (± 0.27)
Beam ($k=5$)	—	—	—	—	—	—	4.47 (± 0.24)	4.23 (± 0.28)
Temperature ($\tau=0.9$)	4.32 (± 0.25)	4.16 (± 0.19)	4.47 (± 0.27)	4.02 (± 0.22)	4.26 (± 0.29)	4.19 (± 0.24)	4.36 (± 0.25)	4.13 (± 0.26)
Temperature ($\tau=1$)	4.36 (± 0.28)	4.25 (± 0.22)	4.47 (± 0.30)	4.02 (± 0.32)	4.2 (± 0.29)	4.18 (± 0.22)	4.42 (± 0.26)	4.15 (± 0.28)
Nucleus ($n=0.9$)	4.32 (± 0.25)	4.28 (± 0.24)	4.48 (± 0.31)	3.99 (± 0.27)	4.16 (± 0.32)	4.13 (± 0.21)	4.39 (± 0.27)	4.13 (± 0.3)
Nucleus ($n=0.95$)	4.3 (± 0.28)	4.28 (± 0.29)	4.49 (± 0.26)	4.00 (± 0.19)	4.24 (± 0.35)	4.14 (± 0.17)	4.44 (± 0.26)	4.08 (± 0.29)
Top- k ($k=30$)	4.35 (± 0.25)	4.21 (± 0.24)	4.53 (± 0.27)	4.03 (± 0.24)	4.2 (± 0.3)	4.16 (± 0.22)	4.44 (± 0.24)	4.18 (± 0.26)
Top- k ($k=40$)	4.34 (± 0.27)	4.24 (± 0.23)	4.53 (± 0.25)	4.00 (± 0.27)	4.17 (± 0.31)	4.11 (± 0.18)	4.41 (± 0.25)	4.17 (± 0.33)
Mirostat ($\tau=3$)	4.39 (± 0.27)	4.26 (± 0.23)	4.55 (± 0.27)	4.02 (± 0.22)	4.16 (± 0.32)	4.17 (± 0.22)	—	—
Typical ($\tau=0.2$)	4.36 (± 0.29)	4.24 (± 0.24)	4.55 (± 0.25)	4.07 (± 0.26)	4.23 (± 0.32)	4.14 (± 0.26)	4.37 (± 0.28)	4.16 (± 0.29)
Typical ($\tau=0.95$)	4.35 (± 0.28)	4.24 (± 0.23)	4.53 (± 0.26)	4.04 (± 0.21)	4.18 (± 0.31)	4.18 (± 0.22)	4.42 (± 0.28)	4.22 (± 0.27)

Table 5: Breakdown of human ratings on quality metrics per task; results for story generation are from fine-tuned versions of GPT-2 medium (m) and large (l). Values in blue are variances.

	Story Generation						
	PPL (g)	PPL (i)	MAUVE (\uparrow)	REP (\downarrow)	Zipf	D (\uparrow)	Human (\uparrow)
Reference	18.28	3.55	—	0.28	1.09	0.85	4.39 (± 0.02)
Temperature ($\tau=0.9$)	60.3 (42.02)	2.93 (-0.62)	0.89	0.24	1.03 (-0.06)	0.88	4.32 (± 0.02)
Temperature ($\tau=1$)	32.64 (14.36)	2.76 (-0.79)	0.95	0.24	1.03 (-0.06)	0.88	4.36 (± 0.02)
Nucleus ($n=0.9$)	9.61 (-8.67)	3.25 (-0.30)	0.9	0.32	1.25 (0.16)	0.83	4.36 (± 0.02)
Nucleus ($n=0.95$)	13.8 (-4.48)	3.78 (0.23)	0.96	0.29	1.17 (0.08)	0.85	4.36 (± 0.02)
Top- k ($k=30$)	7.60 (-10.68)	2.42 (-1.13)	0.90	0.34	1.42 (0.33)	0.8	4.36 (± 0.02)
Top- k ($k=40$)	8.16 (-10.12)	2.77 (-0.78)	0.96	0.33	1.40 (0.31)	0.82	4.37 (± 0.02)
Mirostat ($\tau=3$)	8.17 (-10.11)	3.53 (-0.02)	0.84	0.34	1.32 (0.23)	0.82	4.40 (± 0.02)
Typical ($\tau=0.2$)	17.82 (-0.46)	3.46 (-0.09)	0.88	0.28	1.26 (0.17)	0.85	4.38 (± 0.02)
Typical ($\tau=0.95$)	14.35 (-3.93)	3.78 (0.23)	0.95	0.29	1.17 (0.08)	0.85	4.37 (± 0.02)

Table 6: Automatic quality and diversity metrics, as described in Section 5.1, along with human ratings on the WRITINGPROMPTS dataset. Human ratings are averaged across criteria to form a single metric. Bolded values are the best results among decoding strategies, where for perplexity and Zipf’s coefficient, we take this to be the delta from measurements on human text (numbers in purple). Numbers in blue are standard error estimates. Results are from fine-tuned GPT-2 medium.

Abstractive Summarization (CNN/DailyMail)

Prompt	<p>(CNN) The attorney for a suburban New York cardiologist charged in what authorities say was a failed scheme to have another physician hurt or killed is calling the allegations against his client “completely unsubstantiated.” Appearing Saturday morning on CNN’s “New Day,” Randy Zelin defended his client, Dr. Anthony Moschetto, who faces criminal solicitation, conspiracy, burglary, arson, criminal prescription sale and weapons charges in connection to what prosecutors called a plot to take out a rival doctor on Long Island. “None of anything in this case has any evidentiary value,” Zelin told CNN’s Christi Paul. “It doesn’t matter what anyone says, he is presumed to be innocent.” Moschetto, 54, pleaded not guilty to all charges Wednesday. He was released after posting \$2 million bond and surrendering his passport. Zelin said that his next move is to get Dr. Moschetto back to work. “He’s got patients to see. This man, while he was in a detention cell, the only thing that he cared about were his patients. And amazingly, his patients were flooding the office with calls, making sure that he was OK,” Zelin said. Two other men – identified as James Chmela, 43, and James Kalamaras, 41 – were named as accomplices, according to prosecutors. They pleaded not guilty in Nassau County District Court, according to authorities. Both were released on bail. A requests for comment from an attorney representing Chmela was not returned. It’s unclear whether Kalamaras has retained an attorney. Police officers allegedly discovered approximately 100 weapons at Moschetto’s home, including hand grenades, high-capacity magazines and knives. Many of the weapons were found in a hidden room behind a switch-activated bookshelf, according to prosecutors. The investigation began back in December, when undercover officers began buying heroin and oxycodone pills from Moschetto in what was initially a routine investigation into the sale of prescription drugs, officials said. During the course of the undercover operation, however, Moschetto also sold the officers two semiautomatic assault weapons as well as ammunition, prosecutors said. Moschetto allegedly told officers during one buy that he needed dynamite to “blow up a building.” He later said he no longer needed the dynamite because a friend was setting fire to the building instead. Kalamaras and Chmela are believed to have taken part in the arson, according to prosecutors. “The fire damaged but did not destroy the office of another cardiologist whose relationship with Dr. Moschetto had soured due to a professional dispute,” according to the statement from the district attorney’s office. Moschetto allegedly gave an informant and undercover detective blank prescriptions and cash for the assault and killing of the fellow cardiologist, according to prosecutors. He also requested that the rival’s wife be assaulted if she happened to be present, authorities said. “He was willing to pay \$5,000 to have him beaten and put in a hospital for a few months, and then he said he would pay \$20,000 to have him killed,” said Assistant District Attorney Anne Donnelly, according to CNN affiliate WCBS.</p>
Reference	<p>A lawyer for Dr. Anthony Moschetto says the charges against him are baseless. Moschetto, 54, was arrested for selling drugs and weapons, prosecutors say. Authorities allege Moschetto hired accomplices to burn down the practice of former associate.</p>
Beam	<p>Dr. Anthony Moschetto faces criminal solicitation, conspiracy, burglary, arson and weapons charges. “None of anything in this case has any evidentiary value,” his attorney says.</p>
Temp	<p>Dr. Anthony Moschetto accused of planning to kill doctor, drug dealer. Scheme imploded when a crisis developed in their relationships. Preliminary bond hearing set for August 5.</p>

Nucleus	Dr. Anthony Moschetto, 54, pleaded not guilty to charges Wednesday. Two men – identified as James Chmela and James Kalamaras – were named as accomplices.
Top-<i>k</i>	Dr. Anthony Moschetto is accused of providing police with weapons and prescription drugs. Authorities say he was part of a conspiracy to harm or kill a rival doctor. His attorney calls the allegations against his client “completely unsubstantiated”
Typical	Dr. Anthony Moschetto is charged with crimes including arson, conspiracy, burglary, prescription sale, weapons charges. His attorney says “none of anything in this case has any evidentiary value”
Story Generation (WritingPrompts)	
Prompt	A kid doodling in a math class accidentally creates the world’s first functional magic circle in centuries.

Reference

It was dark and Levi was pretty sure he was lying on his back. There was firelight flickering off of what was left of a ceiling. He could hear something but it was muffled. He was almost positive it was screaming. When he tried to move he felt an excruciating pain in his left side that caused him to cry out. His hand moved to it instinctively and found something protruding from the wound. It seemed to be a pipe of some kind. He attempted to pull it out but it was covered in blood making it slick. Levi's ears began ringing and the sounds began to become clearer, it felt quite a bit like some one was driving needles into his eardrums. The sounds he was hearing were definitely screams and not just one person's, a lot of people were screaming or yelling. There was some one close to him that was crying. He looked in the direction of the tears and seen his sister kneeling beside him, her hands covering her face. "What happened Laur?" He was shocked at the sound that barely rasped out from between his lips. His sister's hands jerked down and she stared down at Levi with a shocked look on her face. "bu... I tho... you weren't breathing!" What started as a whisper ended in yell as she threw her self across her brother and began to sob anew. Levi cried out hoarsely but she didn't hear. She just continued to cling to him like a little girl that had just found her lost doll. He put one of his arms around her and scanned the room as much as he could. It looked like a warzone, like something out of one of the many shooters in his gaming collection. "What the hell?" There were dead bodies everywhere, he recognized some of them. There were firefighters and EMT's running around trying to find survivors in the rubble. Most of the screams were coming from survivors. He seemed to be laying on top of the bottom of a desk, and he was pretty sure the pipe sticking out of his side was a actually one of the legs. Then he spotted it lying about two feet from his right leg, a round section of desk about the size of a softball. On it was a round symbol with geometric shapes glowing with dark red embers and a dull tendril of smoke rising up from it. It all came back to him in rush. He drew that in his notebook. It was second period and his trig teacher had this monotonous voice that could put a crack head to sleep. Laurana caught him doodling and had thrown a pencil at him to scold him silently, which made him jerk as he was connecting the last line on his drawing. Then there was the light and the heat and lastly the dark. Did he do this? What the hell was going on here? A voice brought him out of his little flashback. "Sir. Sir? Sir!" it was a firefighter. "I need you to stay still. You're going to hear a loud noise from behind you. I need you stay calm." Kay. "He just blew up his classroom with a piece of paper and a pencil and this guy wants him to stay calm." it's going to cause this desk to vibrate a lot and your going to feel some pain but there is nothing that we can do about it. We will work as fast as possible." The sound of the saw kicking on was not near as bad as he had thought, but Levi was no where near ready for when it made contact with the leg of the desk. His cries of agony began to drown out the sound of everything around him until darkness overwhelmed him. Sorry about my grammar it's been a while since I have written anything also, first time poster so take it easy on me; p

Nucleus

“Why do you draw a circle?” He asked as he handed the piece of paper over to her. She rolled her eyes and looked down at the picture, “Because I like math!” “Then what?” “I like to draw the circles.” She ran a hand through her curls and looked away. “See, I have three circles.” “Hmph, I didn’t think I saw that in your class.” A younger student brought up the ceiling and murmured. “Yeah, sure. It’s just, why do we draw the circles?” “Because they’re circles.” She looked up. “But they don’t match up!” She pointed at the nearby wall, “They don’t start out like this!” She ran a hand over the map on the wall, “It starts out the same.” She turned and pointed to the map marked with candles. “You have fire here.” “Well that’s nice.” He turned away from her, “I don’t think the artists in class thought about it that way.” She pointed again. “See? Even your circle!” “That’s my circle.” “Look, Mason.” “Hear, daughter.” “I don’t see a man, and so can’t prove it.” The two young adults nodded in agreement. Then she began to speak, “I draw the circles because I like to guess about what will happen when I make circles.” “Yeah, and that’s what we’re going to do when you make a circle and turn the paper to our faces.”

Top-*k*

“Aha. Aah! That makes three.” Professor Gosh sighed. “Oh my God, you had to have been there.” “Well, I’m not sure if that’s the right term, but this circle is going to work.” “But I just told you that it wasn’t working, didn’t you, you little sh**? I’m not kidding! That circle is gon na work!” “So what now?” “Well, that’s for you to figure out. First, you’ve got to figure out a way to communicate with the circle. Maybe the spell isn’t right for it, but it’s not going to ruin the circle.” “Okay, okay, you’ve told me to figure it out so I’ll do a whole lot of work. But how exactly do you communicate a spell to a circle of stones in a basement?” “Simple, really. You put it inside the circle, and it magically works its way around the circle to get to the person you wrote ‘I AM’ on.” “But how? Why am I supposed to be the main caster?” “Because you don’t have magic powers to cast spells.” “What? That’s not even supposed to be possible.” “You are special, you know.” Professor Gosh looked at me with a little bit of sympathy. “I can’t believe you’re a wizard, you can do all this stuff when you’re a kid.” “Well, yeah, that’s what they told me, so I guess you’re right. But no matter how hard you try, you can’t control magic. Don’t worry, we’ll figure out a way to make you a wizard one of your own.” “Okay, I think you understand, I still haven’t figured out how to create a spell, but I am so excited about the possibilities!” “Alright, let’s do this.”

Temp

It was teachers duty to ensure that our kingdom is pure and pure and successful, however it does not mean we do not try to be as rebellious as any other kingdom. Most of our teachers are suitable for rulership, being knowledgeable in sadistic rec thinking of time stopping magic circle created by several apprentices at the same time. After the first amendment, we decided that whilst having a society that thrives on protecting and serving the majority of people, whilst ensuring that the needs of some are met, the elements of freedom, equality and liberty shall not be infringed upon. Sadly, we 've done little to uphold those values since the day our forefathers signed the ancient Papercut act and limited the extent of human freedoms. It is now a day we lay a steady hand on the pages of history, hoping that the rules shall be rewritten and brought to what they once were. The next barrier to seeing the world anew as a black pantheon was a fallen occult student, Mr Pegarchow in Commerce, brought to history via the magic eye how to make a magic circle in the first place. This student made a hoop in the library that can be opened with a simple spiral. If a magic circle is inside this hoop, the magical illusion created is that a circle is made with the appropriate essence of spirit. An image of Pgyarchow being covered by the hoop, with the colours that inhabit it are verily magical. The magician essentially created... magic. To this day, not a single piece of paper hath survived contact with the magic eye. There were so many experiments looking for what function to draw in, or when it starts, 3rd century of my not-so-distant time. Nothing turned out as expected and everyone quickly passed into subculture that embraced the forbidden art of forgetfulness. However, even through the strongest taboo that could cause, symbols of independent spirit were still made and used as religion source material. I am the last wizard right now though not by choice, which was rather problematic after my mentor 's death. However, I am different, and a magician, not by birth but via training and knowledge dedcialized from the single book of knowledge stored in my heart, able to share it with everyone, even those who share their mothers blood. To mhm, Lizonun

ready to go home, when a little kid starts to sketch a circle. “No. Just stop. It’s not working.” “But the circle turns to smoke and then to a flame. It’s not working, I know it is!” Another kid, who was also drawing a circle with his phone, starts to scream. The teacher, looking at the circle, lost his patience and started to yell at the student: “This circle is not working, you need to come up with some more inventive magic demonstration!” “I like it. You can draw it with your phone, it’s not hard.” “No, stop it. Just stop drawing it like that and move on with your day. You look troubled.” “No, I do. I can get some ideas for my magic circle. Also it looks cool and I can tell everyone that it’s working.” “Ok, but you still have to come up with a name for the circle.” “I can come up with kenning, kenning-perforant, kenning-untergrund or kenning-unterhrund depending on the layout. Then my circle is named kenning-unterhrund, and then I post that on reddit, and I sell it on Etsy!” “Uhh... you spelt it wrong again using your phone, kenning-unterhrund for what’s a circle in the middle of a room! You’re not getting any creativity from that!” “Still, I’m not that bad. I could come up with a name too. We could see if our circle works on reddit!” The teacher starts to object, but the kid just brushes him off: “open the door and let me in, I’ll tell everyone my name for my circle.” The door is open and the teacher comes in. “KENNING-UNTER DOES N’T WORK, IT’S BAGA!” The kid answers: “Open the door and let me in. If there’s no door, I’ll come back tomorrow and make a new one.” He sounds confident in his plan. “Ok, kenning-untergrund, kenning-unterhrund or kenning-unteromactually totalitarian need you! Honoring your name will make your circle work better!” The kid is ready to spell it then: “The kenning-untergrund, kenning-unterunterhrund, kenning-unteronAAAKE!” The teacher asks: “What do you mean fails?” “Ok, and I forgot why I just copy all my circles to my phone:)” “You can’t make circles! That would be impossible. And you forgot the reason for your circle to not work:)” “Yes, I did! I forgot the purpose of my circle and this happened to me too:)” “That might happen, but get out of here immediately, the spell is about to hit the ground!” The kid grabs the teacher by the shirt collar: “So what did you get out of it? I thought magic circles are making people happy?” The teacher looks at him with fury in his eyes: “You should go to the yelling and crying section of the school, before the spell hits the ground. You’re going to repeat the ‘Wizard’ spell again.” “Thanks, but no, I’ll leave.” The kid grabs the ground and he leaves the school. Then his circle is gone. The teacher is still trying to figure out what should be done about the circle because he forgot to ask some of his other students. So he goes home to his basement, to his computer and he starts to draw a circle. And then he draws another one. And another one. Soon enough, there are some things that are drawn on his screen: a circle with smoke coming out of it, a circle with fog coming out of it, a circle with fire coming out of it, a circle with flavouring of words coming out of it, a circle with adding words into it and a circle with some mixture of words coming out of it. As the circle is wants to be drawn, he starts to move his finger Mask he was using to draw the circle. When suddenly, his finger don’t move, slowly but steadily, like a hand is starting to move against his will. The students are looking at the screen now. They start to scream and yell: “Why is the hand still moving?!” “Yes,

Typical

As soon as he got the clock on his desk he jumped up. The noise it was making had changed his day forever. His hair, normally unkempt, now glistened and he had to have more hair because of it. "Yes!" He had declared triumphantly as he watched his father fumble around in the backpack in search of a little compass he always kept with him. He took it from the desk and quickly walked out the door to see what had happened. He wasn't the first child in the world to draw something to him and so far, so good. It wasn't as big as a tree, but it did sparkle a little bit and made a few notes in his calculus textbook. As he left the room, the others around him also drew and explained to him their ability. Most kids who knew their way around magic started to teach their spells in a lesson on learning. His teachers tried to tell him about his parents but he only heard his name as they asked if he knew about them. They explained how he would always go with them when he learned something and then asked how he had found the way. His mother said she always did the work when she taught and then how his father used the ingredients he had brought along to create his circle. It wasn't anything like what his mother would use to do her magic but he wasn't interested in her or his magic because it wasn't his favorite type. As he got older he got to watch as the world moved. His family went out on field trips to look at stars and wonder where the other world is. They asked where his mother and father went, and how long they spent on that. It wasn't anything like he expected but they never explained what the field trips were about. As he grew older he became a man with more responsibility and began to notice things like this happen to the rest of the world. As a man who didn't believe in magic he found his circle was missing from the side of the road he always stopped on to check his bearings. His neighbors thought it was weird that his neighbors' circles would get in the way. One day his friends stopped on the street he was walking down and told him to get to his car and take him to a store so they could find his circle. When they found his circle they gave it back and it had no clue where it came from. After some further digging he discovered the missing part of the circle. A pentagram that appeared out of nothingness on the road and for the longest time, no one had been able to find it. A couple weeks later a truck came out and dumped all of the construction workers that were using the road and blocked it. When his circle had disappeared the news talked about the weird pentagram on the road. This went on for months until his mother noticed he wasn't coming home. When she came home, she looked in his room and he wasn't there. When he wasn't there he never made his presence known to her and never tried to teach his spells to her. They eventually went back to their apartment together. Her father didn't even acknowledge his daughter at first but his wife didn't think to call him or get his attention. The father, seeing the pattern and a good time with his daughter decided to stay. She told her parents to check their calendar. It had stopped being so late that her father thought to take his place to try to sleep off his spell exhaustion. The parents realized something was up and the wife suggested she get her boyfriend and see if she could have a quiet time. It wasn't a normal time to do such things, but when he found her and saw his father had passed he told his girlfriend that they would meet for drinks to get over it. The wife got ready to go, and they drove home. She pulled over in front of their house. A truck pulled up in the driveway, stopped and looked out and began to slowly pull into the driveway. When she was looking in his eyes, she asked if he knew about her mother's strange powers. When he replied he asked how his father got the circle that they kept around for all of these years.

The man pulled up in his truck and gave his girlfriend the back seat and started the truck. She took him home, they drank, and watched a movie about dragons and aliens and made up a bedtime story for the boy and girl. After dinner the boy slept through the night. He dreamt that the circle had gone missing, that it had disappeared in his yard, and that the car parked next to him was covered in dust and it smelled like his room had been raided by monsters. After his sleep, the mother took him to his room to wake him up. When she entered he didn't react at all. "Are you alright honey?" The mother asked him. The boy responded in the most puzzled tone she had ever heard him.

Table 7: Full sample generations for abstractive summarization and story generation. We use samples from the model fine-tuned from GPT-2 large for story generation.